Course #412
Analyzing Microarray Data using the mAdb System

September 14-15, 2004 1:00 pm - 4:00pm

madb-support@bimas.cit.nih.gov

- Intended for users of the mAdb system who are familiar with mAdb basics
- Focus on analysis of multiple array experiments

Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
   – Class Discovery - clustering, PCA, MDS
   – Class Comparison - statistical analysis
   – Class Prediction - PAM

Various Hands-on exercises

1. mAdb system overview

mAdb Data Workflow

Upload Data

File Format

- GenePix
- MAS5
- GCOS 1.1
- ArraySuite

Quality Control

Project Summary

- Summary Statistics
- Array images
- Graphical Report

Prepare Dataset

Dataset Extraction

- Normalization
- Spot Filtering

Analysis/Model

Analysis Tools

- Class Discovery
- Class Comparison
- Class Prediction

Review Annotation

Annotation Tools

- Feature Report
- Gene Ontology
- BioCarta Pathway
- KEGG Pathway

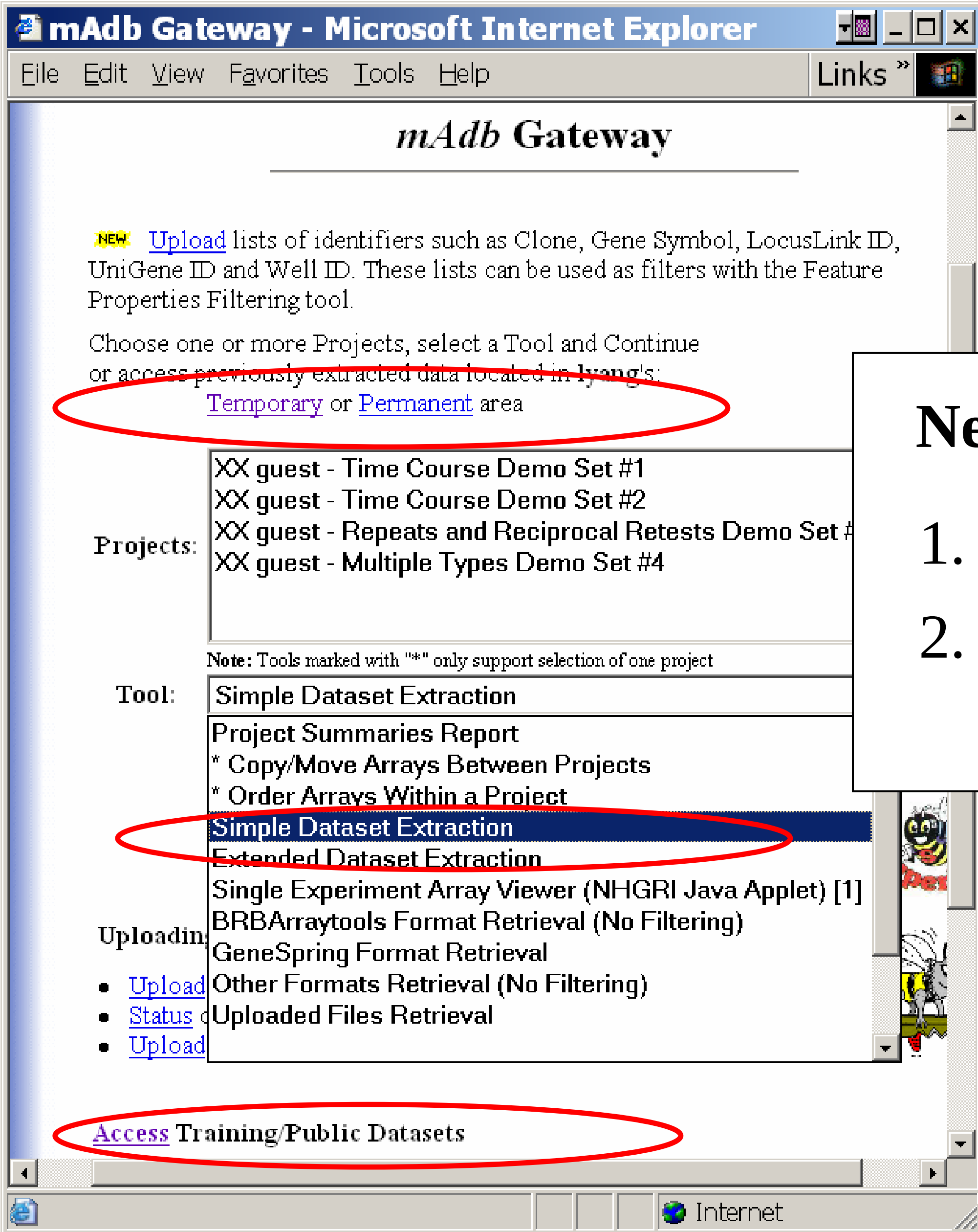
2. mAdb dataset overview

What is a dataset?

- mAdb Dataset
  - Collection of data from multiple experiments
  - Genes as rows and experiments as columns

Genes		sample1	sample2	sample3	sample4	sample5	…
	1	0.46	0.30	0.80	1.51	0.90	...
	2	-0.10	0.49	0.24	0.06	0.46	...
	3	0.15	0.74	0.04	0.10	0.20	...
	4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	5	-0.06	1.06	1.35	1.09	-1.09	...

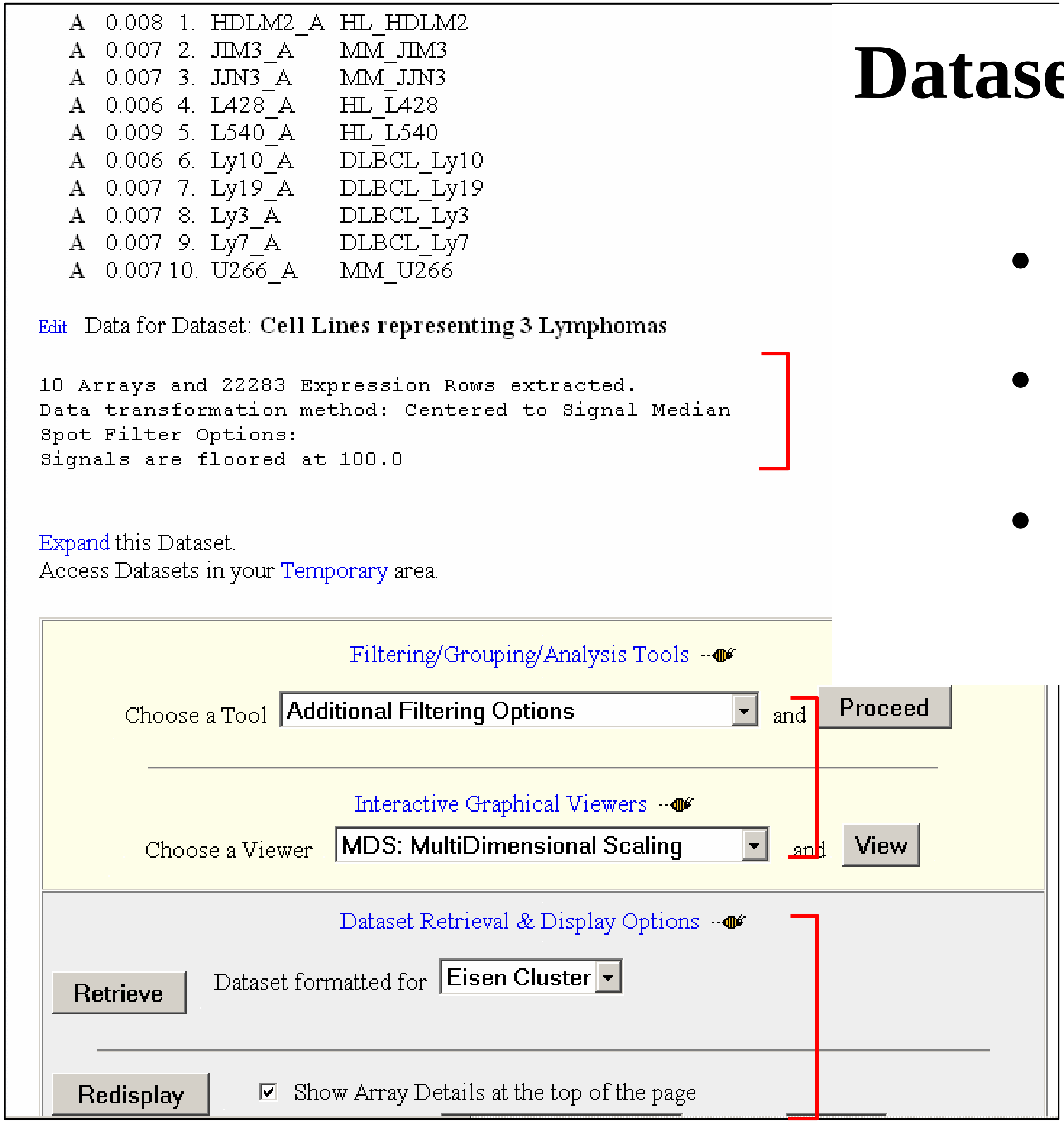
Gene expression level **=** (normalized) Log(Red signal / Green signal)



New or Existing Dataset:

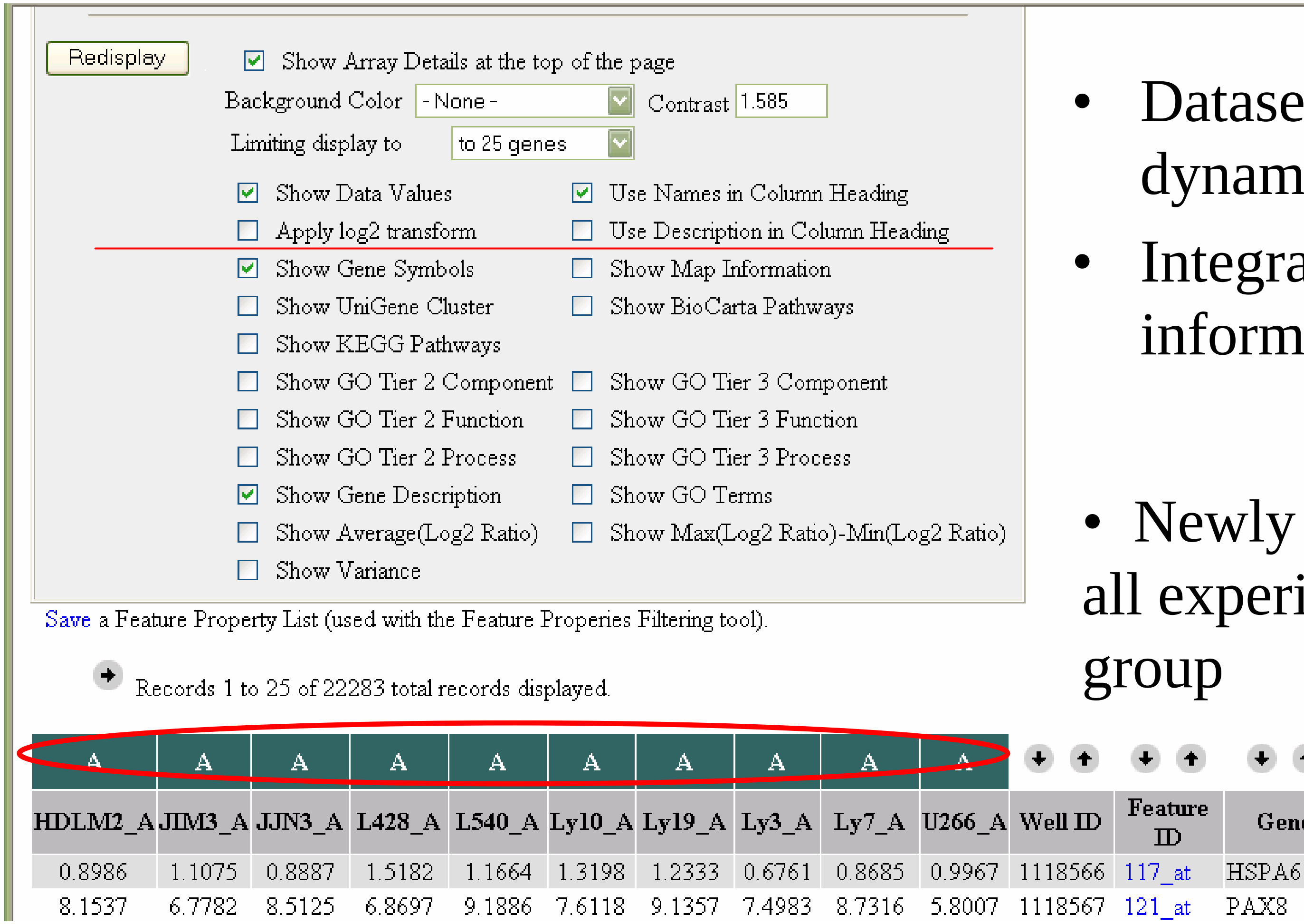
1. Create New Dataset
2. Access Existing Dataset

Dataset Display Page



- Dataset History
- Analysis Tools
- Retrieval and Display Options…

Dataset Display



- Dataset display options dynamic
- Integrated gene information

- Newly created dataset puts all experiments into a single group

mAdb Dataset Display

Group label

Sample name

genes

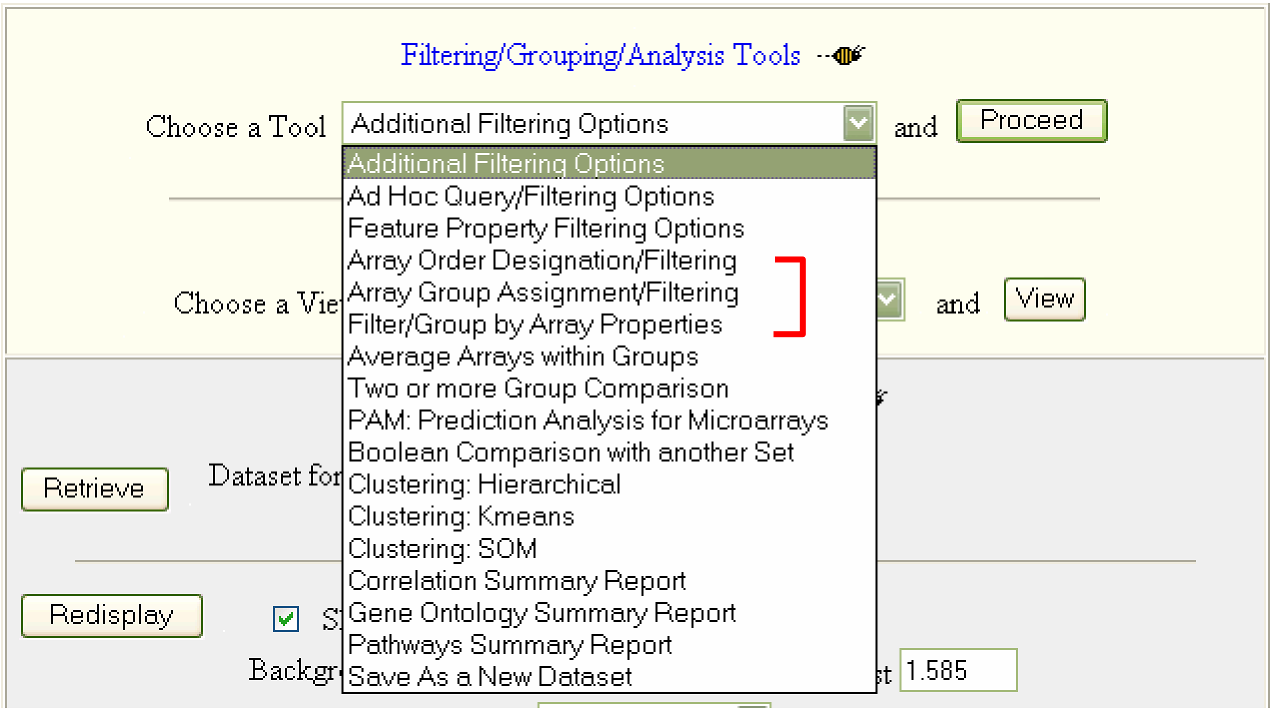
Group label	A	A	A	A	A
Sample name	BJAB_A_B	Daudi_A_B	Jurkat_A_B	Ly10_A_B	Ly3_A_B
				7.7702	
	9.7305	9.7985	9.7249	10.2981	10.1150
		8.9715			
		8.8918	9.0752	10.2200	
	8.4250	7.0224	7.8511	7.4692	7.7886
	6.9189	7.5645			7.7814
	9.3296	9.6202	9.4409	9.9652	10.0534
				7.8629	7.3505
	10.0053	9.6605	9.3872	9.9003	9.3181
	8.1908	8.2187	7.3540	8.3650	
	6.5014			7.0629	
		6.5251	6.4512		
	9.6604	10.0402	8.6991	9.9747	9.4539
	8.3781	8.8981	8.1739	8.2322	9.3807
	7.9419	7.4741	7.9301		
	8.9372	9.8243	9.4774	9.7465	10.2738
	8.2002			9.9105	9.6255
	5.0575	6.8163	5.9542		5.7388
	9.9564	9.8420	9.7677	10.1529	9.3419
	9.9284	9.6363	9.3726	9.8858	10.1808
	9.4419	9.0507	9.4075	9.9434	9.0739
	10.4035	9.7502	9.2389	10.1029	10.5434
	9.0906	9.3452	9.3869	9.6770	9.3613

Well ID	Feature ID	Gene	Description
1118566	117_at	HSPA6	heat shock 70kDa protein 6 (HSP70B')
1118567	121_at	PAX8	paired box gene 8
1118568	177_at	PLD1	phospholipase D1, phophatidylcholine-sp
1118569	179_at	PMS2L9	postmeiotic segregation increased 2-like
1118570	320_at	PEX6	peroxisomal biogenesis factor 6
1118572	564_at	GNA11	guanine nucleotide binding protein (G pro
1118573	632_at	GSK3A	glycogen synthase kinase 3 alpha
1118574	823_at	CX3CL1	chemokine (C-X3-C motif) ligand 1
1118575	1053_at	RFC2	replication factor C (activator 1) 2, 40kD
1118576	1294_at	UBE1L	ubiquitin-activating enzyme E1-like
1118577	1316_at	THRA	thyroid hormone receptor, alpha (erythro
1118579	1431_at	CYP2E1	cytochrome P450, family 2, subfamily E
1118581	1487_at	ESRRA	estrogen-related receptor alpha
1118582	1729_at	TRADD	TNFRSF1A-associated via death domain
1118584	1861_at	BAD	BCL2-antagonist of cell death
1118585	243_g_at	MAP4	microtubule-associated protein 4
1118586	266_s_at	CD24	CD24 antigen (small cell lung carcinoma
1118587	31799_at		Sapiens clone 24627 mRNA sequence
1118588	31807_at	DDX49	DEAD (Asp-Glu-Ala-Asp) box polypepti
1118589	31826_at	KIAA0674	KIAA0674 protein
1118591	31837_at	BC002942	hypothetical protein BC002942
1118592	31845_at	ELF4	E74-like factor 4 (ets domain transcripti
1118594	31861_at	IGHMBP2	immunoglobulin mu binding protein 2

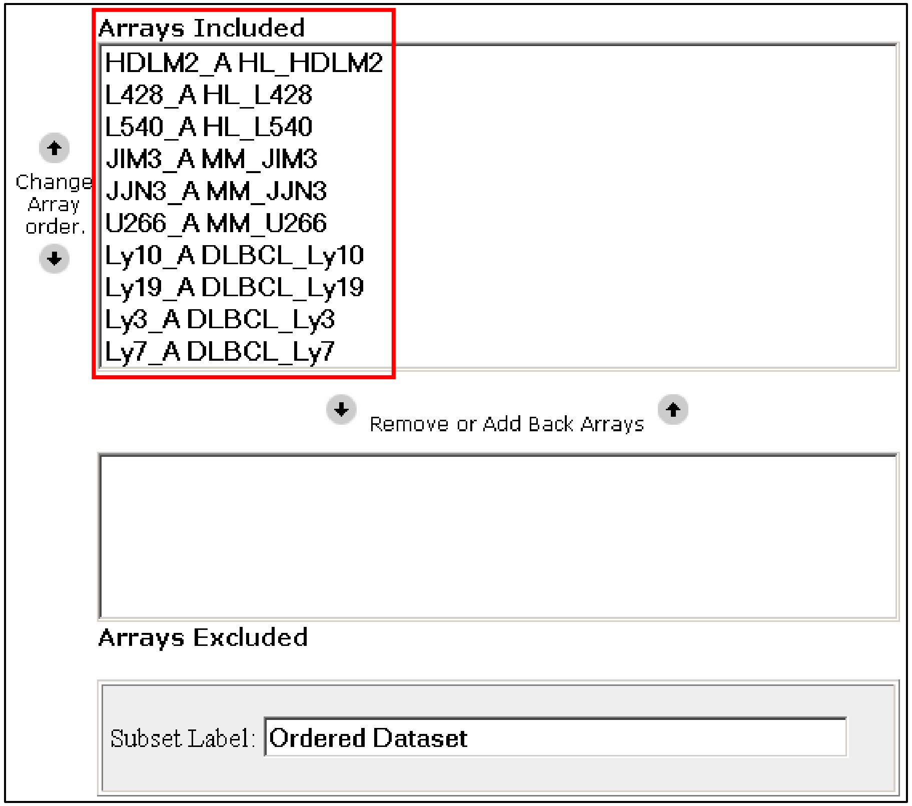
Dataset Group Assignment

- Array Order Designation/Filtering
- Array Group Assignment/Filtering
- Filter/Group by Array Properties

Dataset group assignment tools

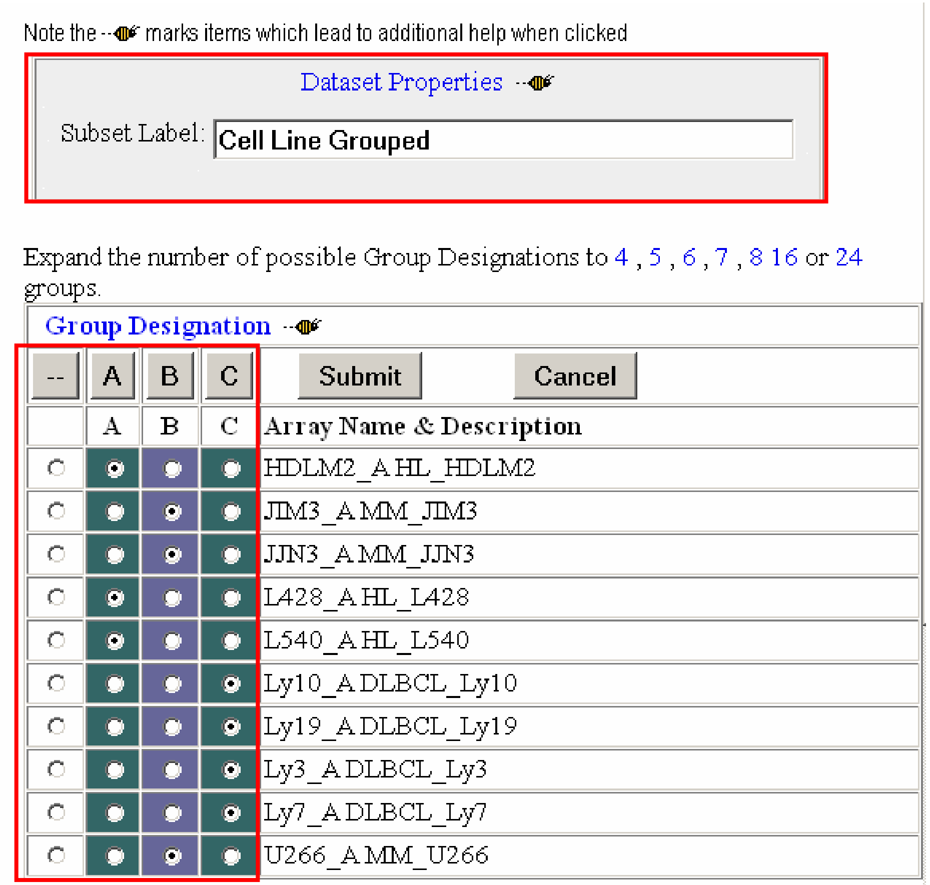


Array Order Designation/Filtering



- Order arrays in dataset
- Delete/Add back arrays in dataset
- Subsequent analysis will be ordered by groups first and then ordered within each group
- Does not group arrays

Array Group Assignment/Filtering



- One click per array for additional group
- Not convenient for large dataset
- Can not order within group

Filter/Group by Array Properties

mAdb Dataset Display

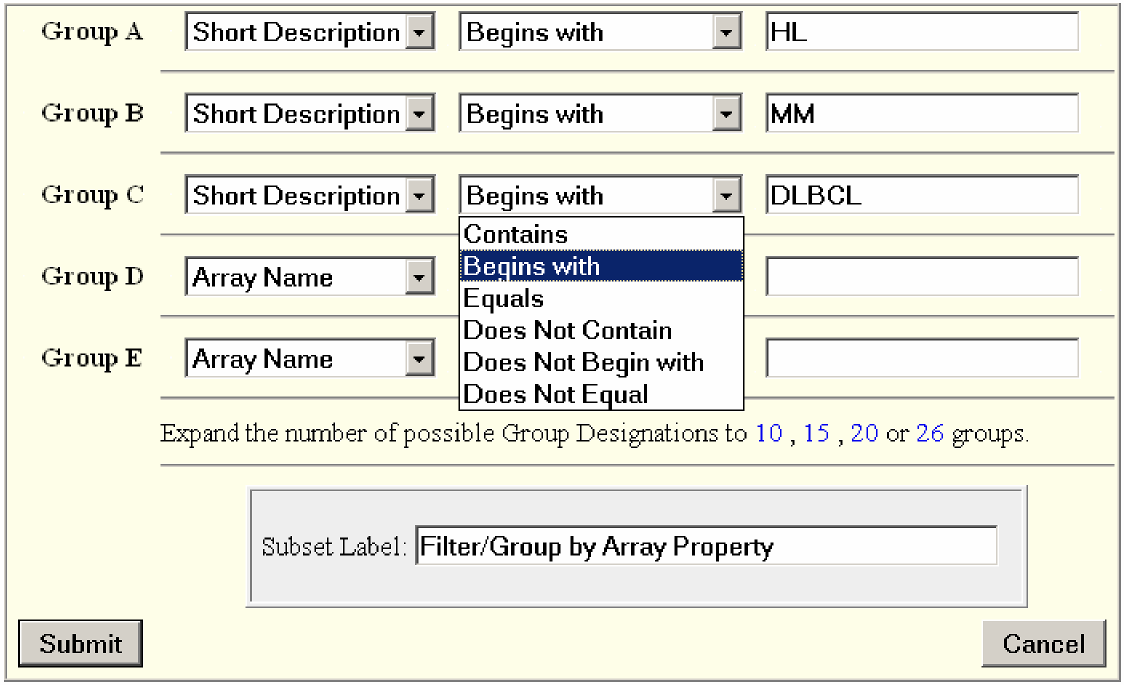
A	0.008	1.	HDLM2_A	HL_HDLM2
A	0.007	2.	JIM3_A	MM_JIM3
A	0.007	3.	JJN3_A	MM_JJN3
A	0.006	4.	L428_A	HL_L428
A	0.009	5.	L540_A	HL_L540
A	0.006	6.	Ly10_A	DLBCL_Ly10
A	0.007	7.	Ly19_A	DLBCL_Ly19
A	0.007	8.	Ly3_A	DLBCL_Ly3
A	0.007	9.	Ly7_A	DLBCL_Ly7
A	0.007	10.	U266_A	MM_U266

Edit Data for Dataset: **Cell Lines representing 3 Lymphomas**

```
10 Arrays and 22283 Expression Rows extracted.
Data transformation method: Centered to Signal Median
Spot Filter Options:
Signals are floored at 100.0
```

- Array properties include Name and Short Description
- Identify consistent pattern

Filter/Group by Array Properties



- Convenient for large dataset
- Can not order arrays within group

Group Assignment

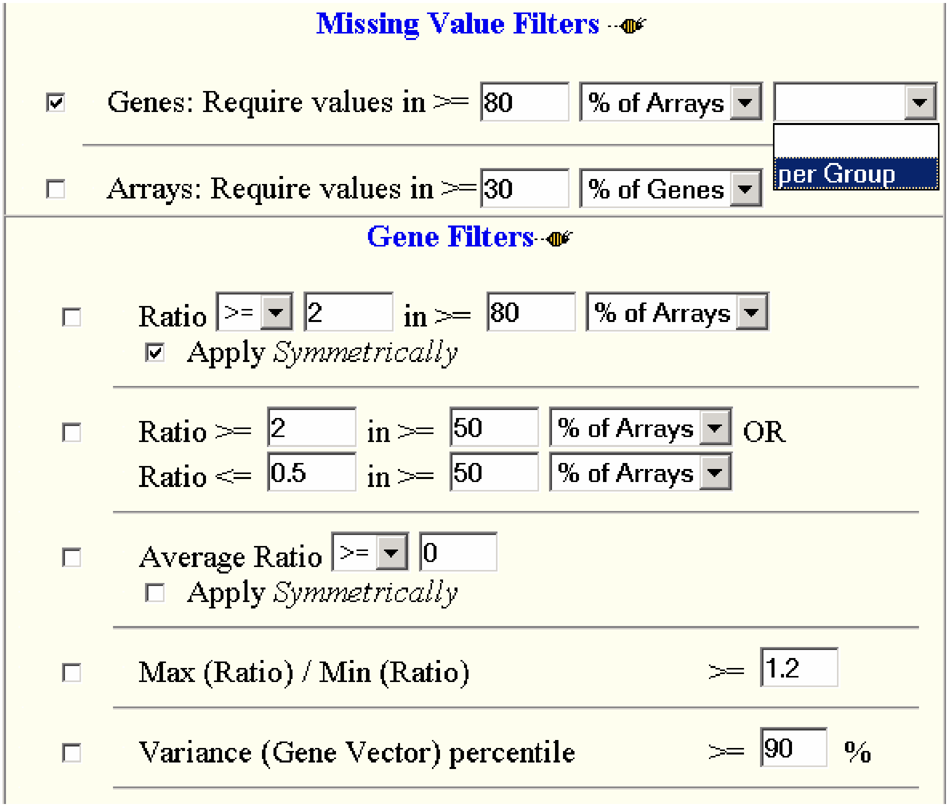
A	A	A	B	B	B	C	C	C	C			
HDLM2_A	L428_A	L540_A	JIM3_A	JJN3_A	U266_A	Ly3_A	Ly7_A	Ly10_A	Ly19_A	Well ID	Feature ID	Gene
0.8986	1.5182	1.1664	1.1075	0.8887	0.9967	0.6761	0.8685	1.3198	1.2333	1118566	117_at	HSPA6
8.1537	6.8697	9.1886	6.7782	8.5125	5.8007	7.4983	8.7316	7.6118	9.1357	1118567	121_at	PAX8
0.8042	2.2147	0.8831	0.6680	0.6954	1.4118	0.6761	0.6743	0.6046	0.7337	1118568	177_at	PLD1
4.1856	6.4728	9.8080	5.3601	6.0779	5.1954	7.1981	3.7505	7.2110	4.8481	1118569	179_at	PMS2L9
2.3557	1.6427	1.2628	2.5865	2.4068	2.0954	1.4949	2.1160	1.0713	2.5561	1118570	320_at	PEX6
1.1856	1.3852	0.9514	0.9599	0.9757	0.8588	1.2529	1.4626	1.3452	1.2318	1118571	336_at	TBXA2R
3.7746	1.6271	2.5043	1.1516	1.0508	0.6536	1.4875	1.9670	1.1227	1.1988	1118572	564_at	GNA11
4.5008	5.1783	5.5333	5.3079	7.4172	6.8863	7.1846	5.8658	6.0435	8.4519	1118573	632_at	GSK3A
4.1646	12.1329	0.8532	0.6680	0.6954	0.6536	1.1034	0.6743	1.4075	0.7337	1118574	823_at	CX3CL1
5.5663	4.3223	5.4480	1.6206	2.9270	4.4418	4.3158	3.3790	5.7775	3.3067	1118575	1053_at	RFC2
3.9173	2.4157	2.0461	1.3460	0.9437	1.1039	1.3083	2.0964	1.9933	1.9391	1118576	1294_at	UBE1L
0.7800	0.7918	0.8532	0.7715	0.6954	0.8327	0.6761	0.8483	0.8083	0.7630	1118577	1316_at	THRA
0.7800	0.6485	0.8532	0.6680	0.6954	0.6536	0.6761	0.6743	0.6046	0.7337	1118578	1320_at	PTPN21

- Group assignment information is carried into relevant analysis
- Dataset is independent from microarray platforms

Examples for using group labels

- Additional Filtering per Group
- Correlation Summary Report
- Average Arrays within Groups

Filter by Group Properties



- Ensures each group has sufficient number of non-missing values

Correlation Summary Report

Correlations

A	A	A	B	B	B	C	C	C	C				
#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	Grp		Array Name	Array Description
#1 A	0.890	0.914	0.844	0.873	0.852	0.853	0.838	0.856	0.836	A	1.	HDLM2_A	HL_HDLM2
	#2 A	0.882	0.852	0.860	0.847	0.856	0.824	0.869	0.845	A	2.	L428_A	HL_L428
		#3 A	0.860	0.880	0.855	0.858	0.850	0.859	0.843	A	3.	L540_A	HL_L540
			#4 B	0.896	0.895	0.852	0.826	0.850	0.846	B	4.	JIM3_A	MM_JIM3
				#5 B	0.885	0.868	0.853	0.859	0.867	B	5.	JJN3_A	MM_JJN3
					#6 B	0.857	0.832	0.852	0.848	B	6.	U266_A	MM_U266
						#7 C	0.871	0.924	0.882	C	7.	Ly10_A	DLBCL_Ly10
							#8 C	0.873	0.918	C	8.	Ly19_A	DLBCL_Ly19
								#9 C	0.883	C	9.	Ly3_A	DLBCL_Ly3
									#10 C	C	10.	Ly7_A	DLBCL_Ly7

- Pair wise correlation between 2 samples in dataset
- Individual scatter plot available
- Group pattern for quality control

mAdb Correlation Report - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Search Favorites Media

Address http://madb.nci.nih.gov/cgi-bin/restricted/mAdb_correlation_report.cgi Go Links

Visual Bivariate Data Analysis

Home Page | mAdb Gateway | Upload Status
Forums | Reference Info | Program Downloads | GeneCards

mAdb Correlation Report

View Array Summaries

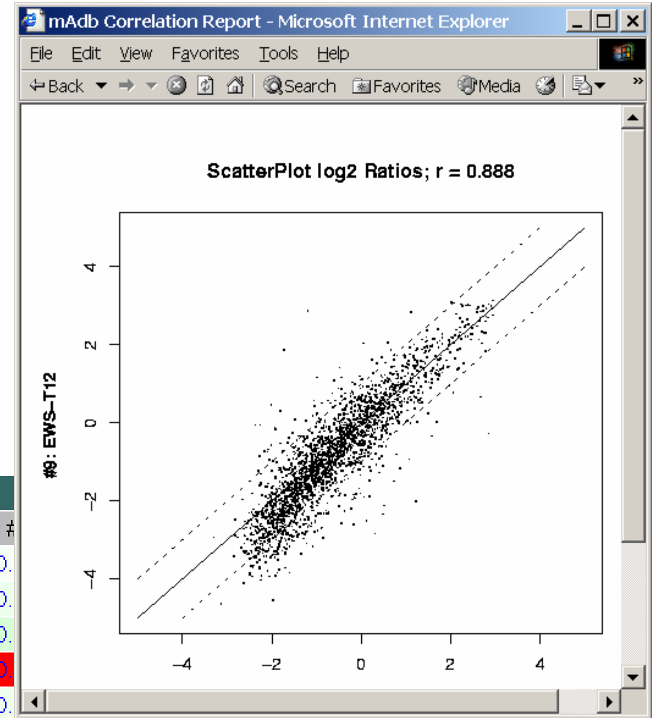
Return to the input dataset.

Redisplay Background Color Scheme Green/White/Red

Color Saturation Max/Mid/Min 0.8 0.6 0.4

Note: For proper coloring Max > Mid > Min

Note: Click on the Correlation values to display the corresponding ScatterPlot

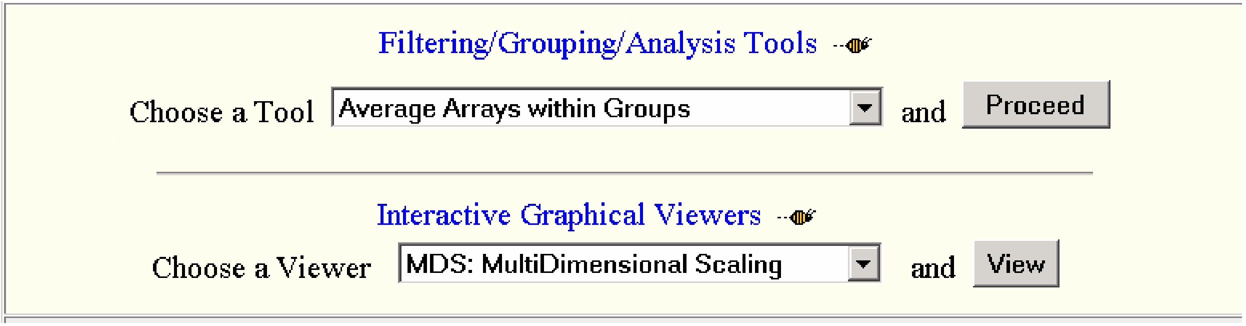


Correlations

	A	A	A	A	A	A	A	A	A	A	A	A									A
	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	#11	#12	#								#21
1.	#1 A	0.809	0.815	0.596	0.618	0.711	0.653	0.723	0.743	0.506	0.682	0.693	0.								.732
2.		#2 A	0.820	0.583	0.528	0.602	0.603	0.645	0.659	0.395	0.549	0.597	0.								.767
3.			#3 A	0.615	0.626	0.689	0.651	0.719	0.699	0.562	0.652	0.659	0.								.673
4.				#4 A	0.497	0.468	0.442	0.513	0.507	0.385	0.432	0.518	0								.517
5.					#5 A	0.784	0.665	0.711	0.731	0.681	0.733	0.726	0.								.492
6.						#6 A	0.727	0.804	0.802	0.659	0.831	0.807	0.730	0.523	0.548	0.501	0.609	0.635	0.641	0.629	0.551
7.							#7 A	0.756	0.751	0.639	0.764	0.799	0.690	0.632	0.560	0.516	0.511	0.636	0.646	0.643	0.523
8.								#8 A	0.888	0.638	0.803	0.808	0.787	0.563	0.588	0.546	0.561	0.633	0.643	0.615	0.573
9.									#9 A	0.623	0.815	0.828	0.725	0.604	0.639	0.623	0.591	0.664	0.672	0.660	0.629
10.										#10 A	0.712	0.679	0.608	0.444	0.330	0.261	0.344	0.409	0.423	0.453	0.290
11.											#11 A	0.881	0.761	0.591	0.519	0.475	0.535	0.646	0.656	0.694	0.496
12.												#12 A	0.711	0.639	0.562	0.533	0.534	0.668	0.676	0.698	0.538
13.													#13 A	0.476	0.466	0.413	0.570	0.588	0.598	0.580	0.529
14.														#14 A	0.552	0.530	0.418	0.799	0.807	0.777	0.497
15.															#15 A	0.891	0.625	0.663	0.668	0.641	0.796
16.																#16 A	0.641	0.642	0.645	0.606	0.821

http://madb.nci.nih.gov/cgi-bin/restricted/mAdb_correlation_report.cgi#main_options_table Internet

Average Arrays within Group



- Averages using log ratios - though user chooses to display linear or log2 values

Dataset I
Small Round Blue Cell Tumors (SRBCTs)

- Khan et al. *Nature Medicine* 2001
- 4 tumor classifications
- 63 training samples, 25 testing samples, 2308 genes
- Neural network approach

Hands-on Session 1

- Lab 1- Lab 4
- Read the questions before starting, then answer them in the lab.
- Use web site: http://mAdb-training.cit.nih.gov.
- Avoid maximizing web browser to full screen.
- Total time: 15 minutes

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Analysis Overview

Class Discovery - Unsupervised	• Clustering – Hierarchical, K-means, SOMs • Principal components Analysis (PCA) • Multidimensional Scaling (MDS)
Class Comparison - Supervised	• paired t-tests • t-test pooled (equal) variance • t-test separate (unequal) variance • Wilcoxon Rank-Sum (Mann Whitney U) • Wilcoxon Matched-pairs Signed Rank • One way ANOVA • Kruskal-Wallis
Class Prediction - Supervised	Prediction Analysis for Microarrays (PAM)

Class Discovery Example

- Discover cancer subtypes by gene expression profiles
- Identify genes which have different expression patterns in different groups

- Tools: PCA, MDS, and Cluster Analysis

Class Comparisons Example

- Find genes which are differentially expressed among cancer groups
- Find genes up/down regulated by drug treatment

- Tools:
  - Two or more group comparison
  - Statistics Results filtering

Class Prediction Example

- Identify an expression profile which correlates with survival in certain cancers
- Identify an expression profile which can be used to diagnose different types of lymphomas

- Tools: Prediction Analysis for Microarrays (PAM)

3. mAdb dataset analysis tools

- Class Discovery: clustering, PCA, MDS
- Class Comparison: statistical analysis
- Class Prediction: PAM

Class Discovery

- Dataset with large amount of data
- Dataset not organized
- Visualization with Clustering, PCA, MDS

Cluster Analysis

- Organize large microarray dataset into meaningful structures
- Visualize and extract expression patterns

What to Cluster?

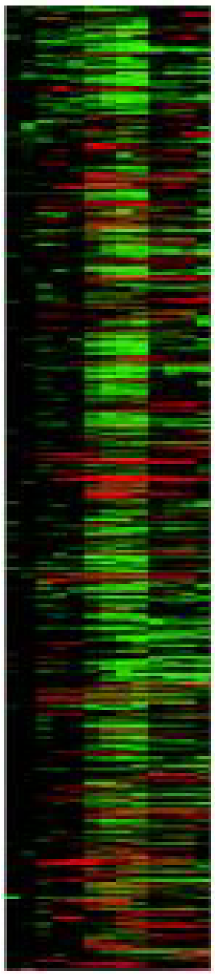
Genes - identify groups of genes that have correlated expression profiles

Samples - put samples into groups with similar overall gene expression profiles

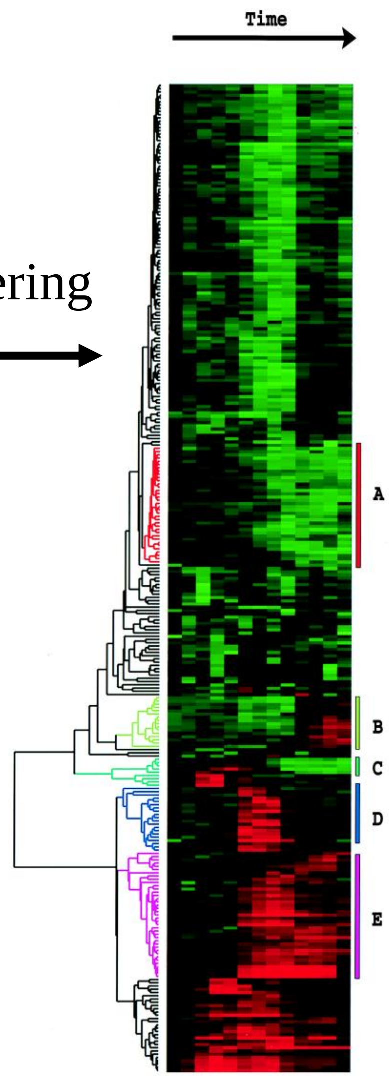
Clustering Methods

- Hierarchical clustering
- Partitional clustering
  - K-means
  - Self-Organizing Maps (SOM)

Cluster Example on Genes



Clustering



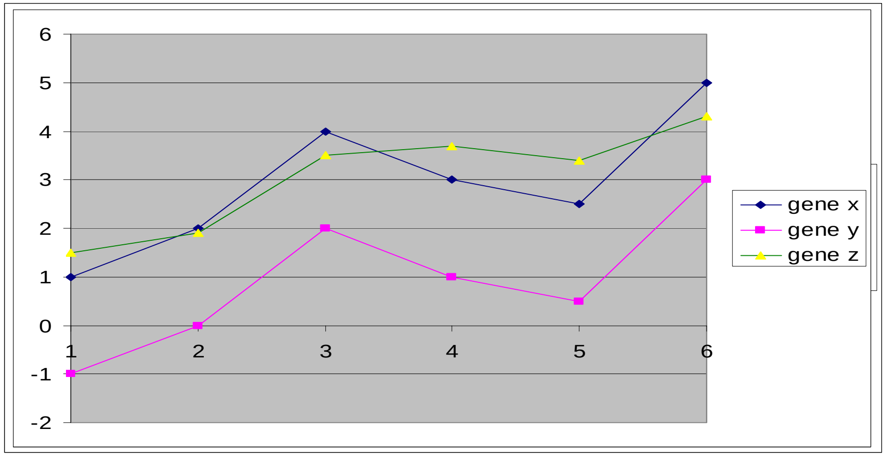
Much easier to look at large blocks of similarly expressed genes

Dendogram helps show how 'closely related' expression patterns are

A. Cholesterol syn.
B. Cell cycle
C. Immediate-early response
D. Signaling
E. Tissue remodeling

2 Steps

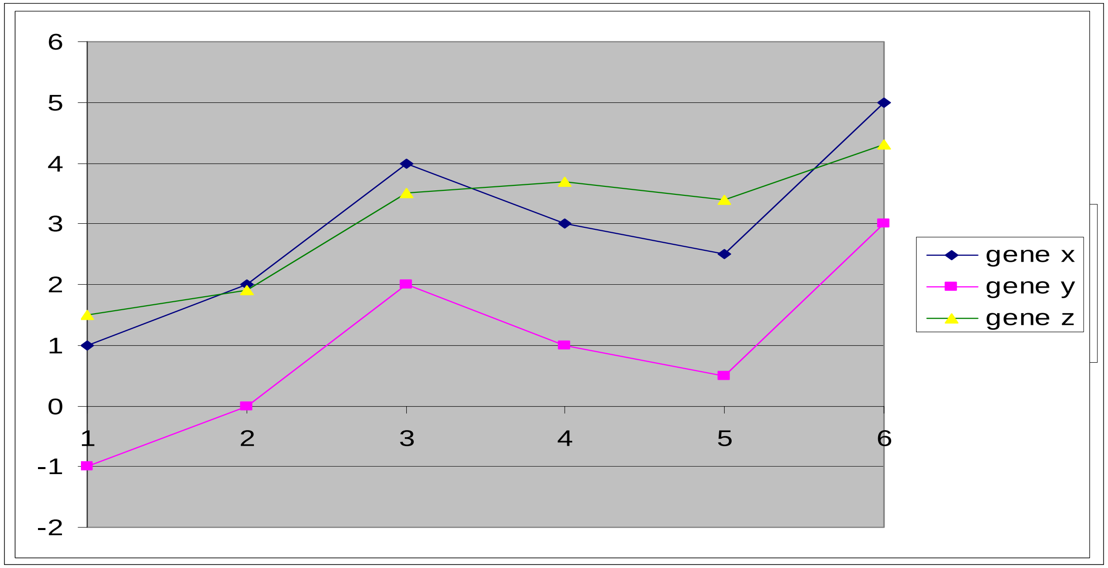
- Pick a distance method
  - Correlation
  - Euclidian



- Pick the linkage method
  - Average linkage
  - Complete linkage
  - Single linkage

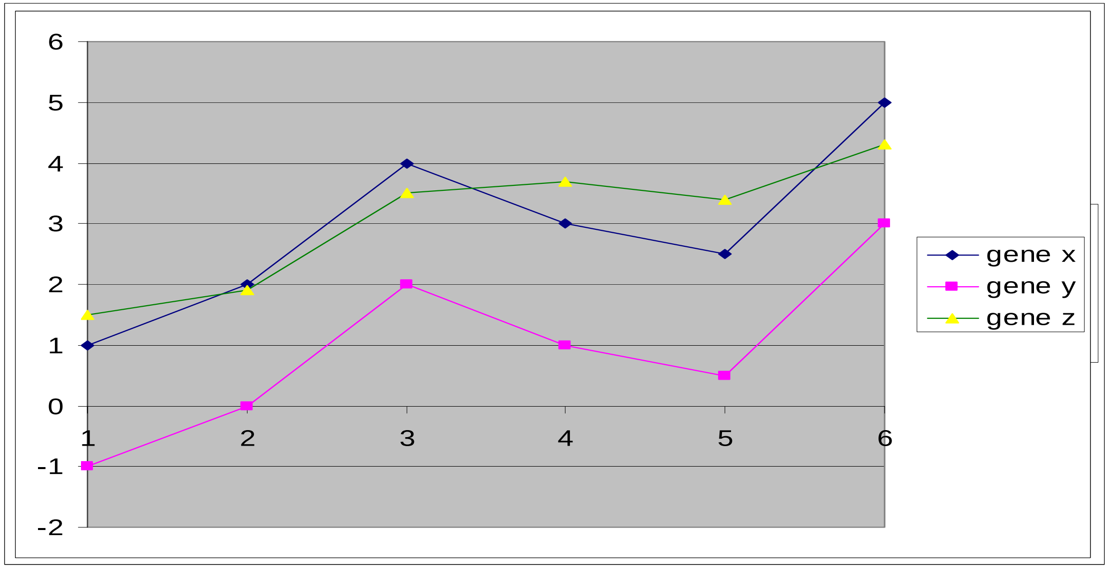
Correlation

- Compares shape of expression curves (-1 to 1)
- Can detect inverse relationships (absolute correlation)

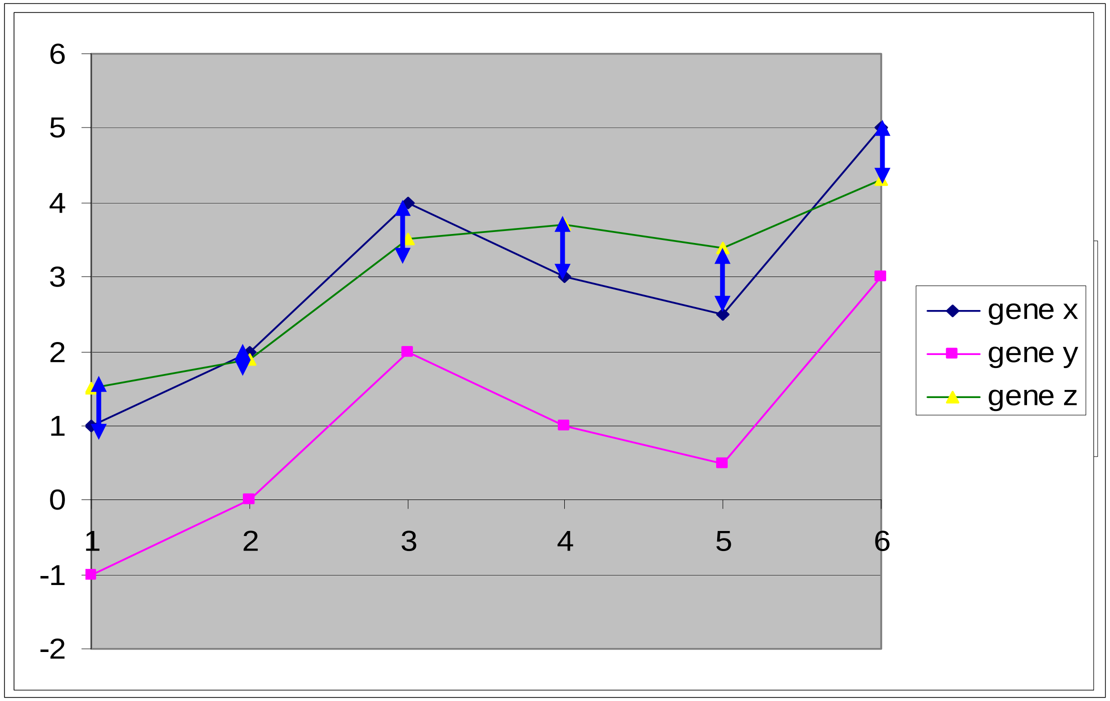


Two Flavors of correlation

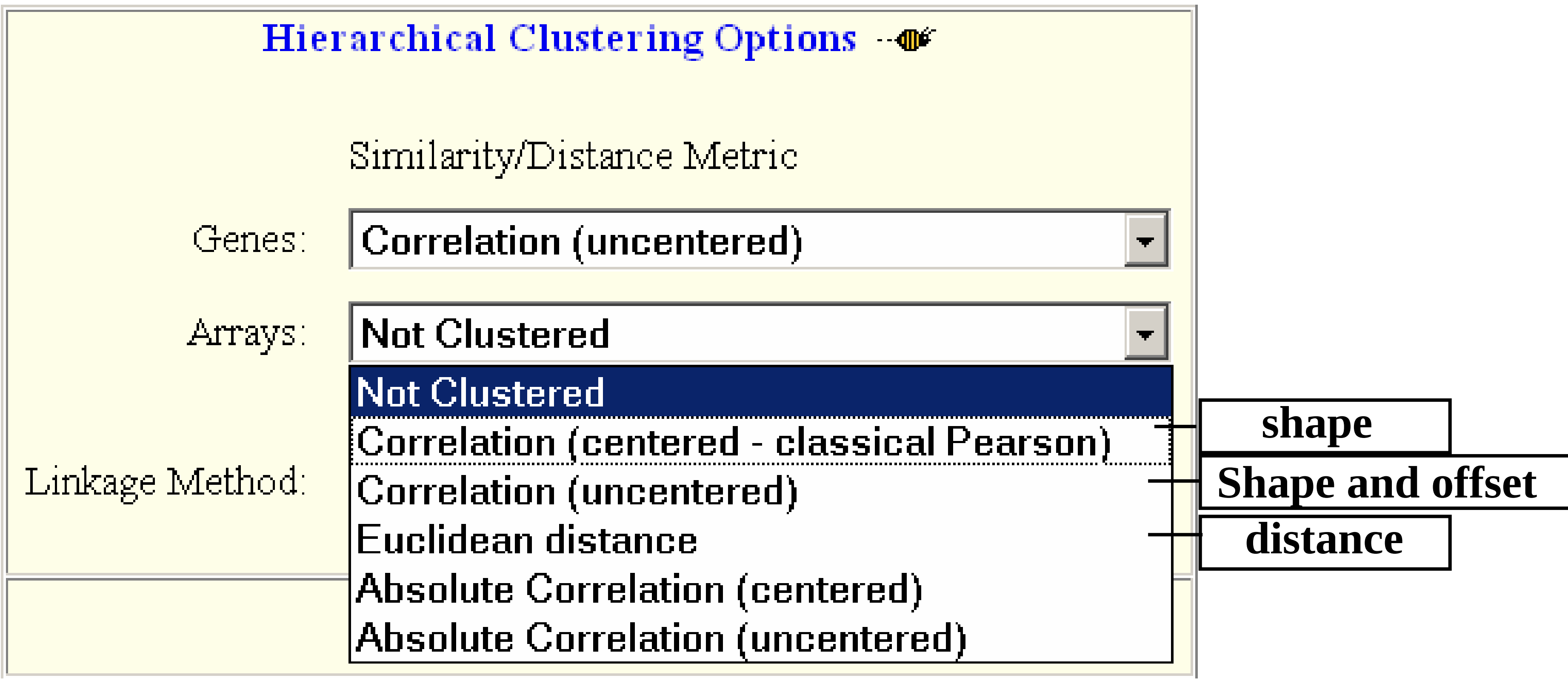
- Correlation (centered-classical Pearson)
- Correlation (un-centered)
  - assume the mean of the data is 0, penalize if not
  - Measures both similarity of shape and the offset from 0



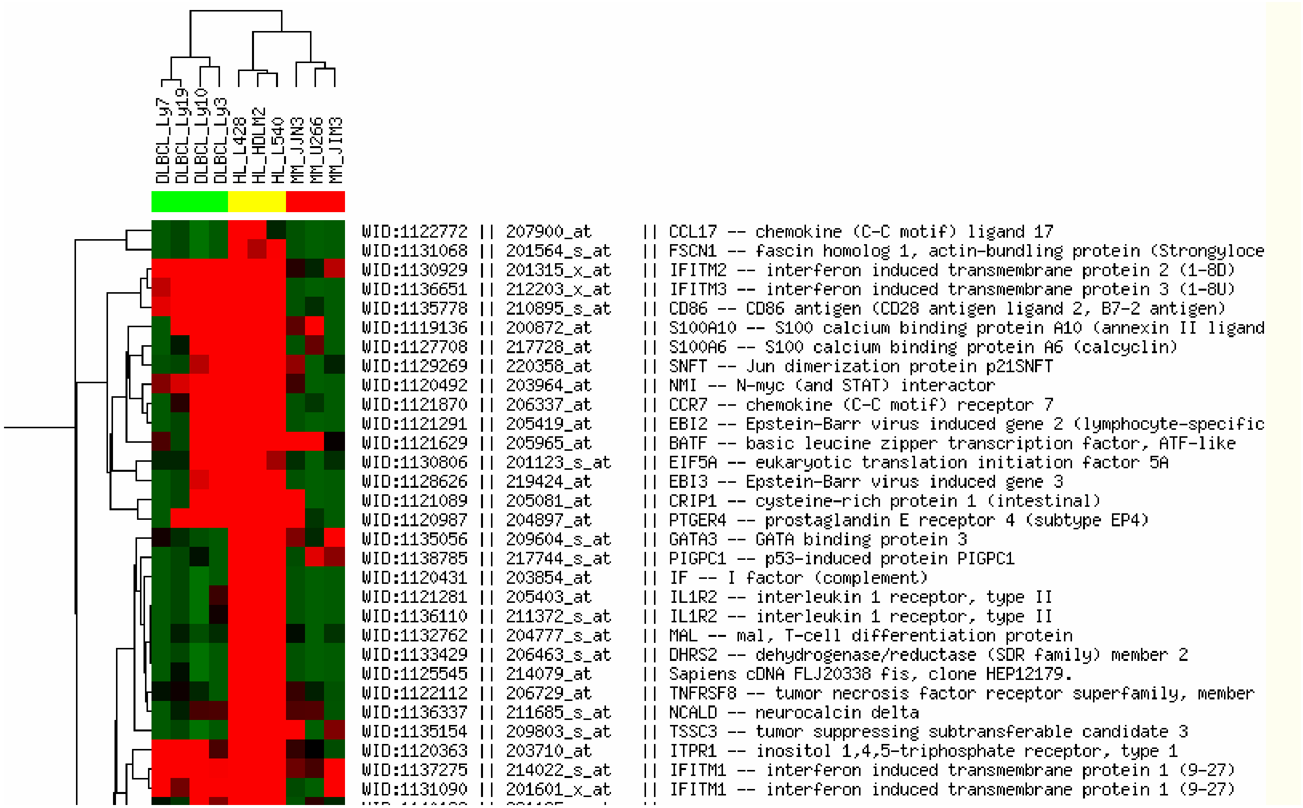
Euclidean Distance



Similarity/Distance Metric Summary

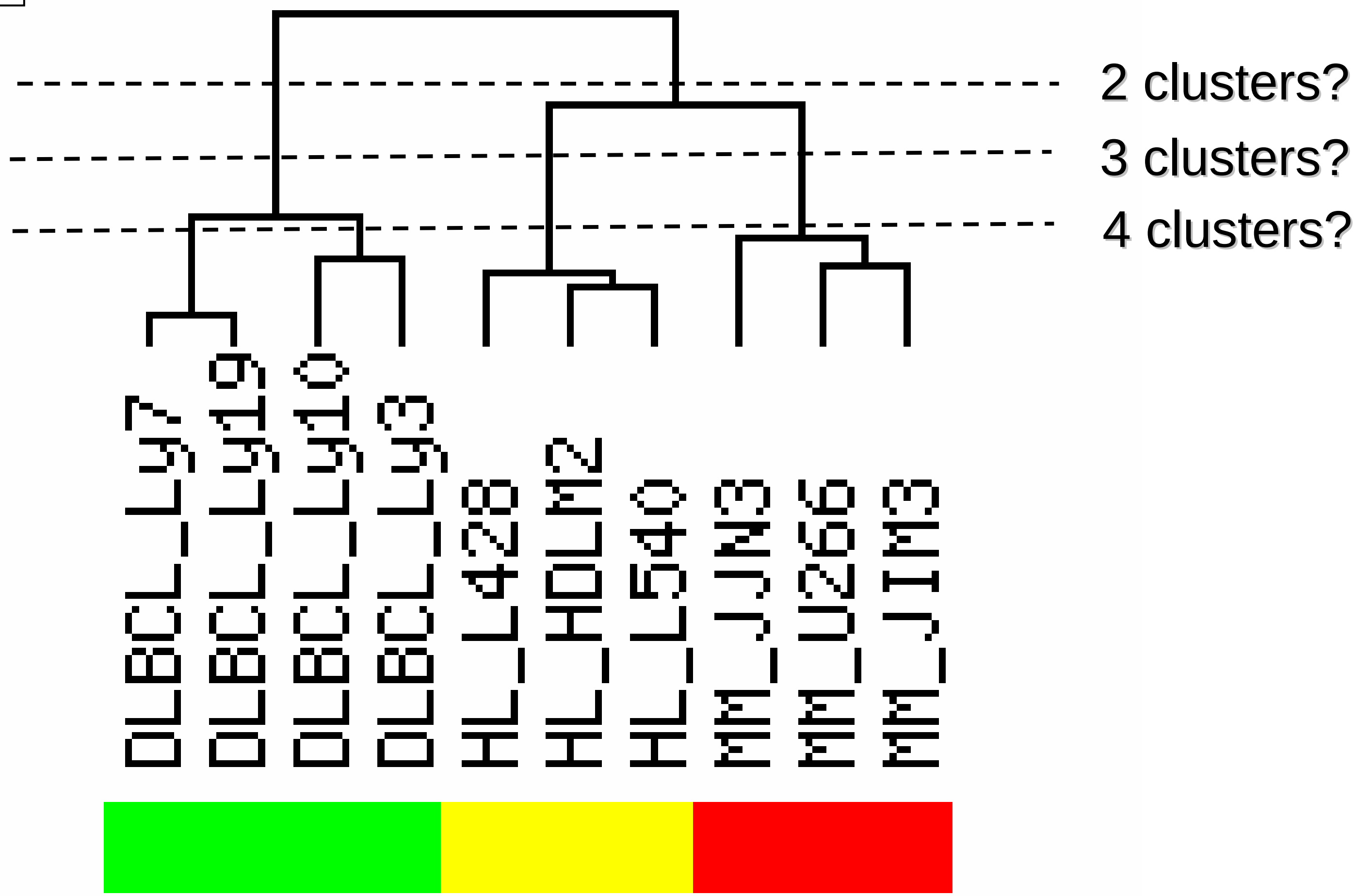


Hierarchical Clustering Example



Tree Cutting

Degrees of dissimilarity



Hierarchical Clustering Summary

- Detection of patterns for both genes and samples
- Good visualization with tree graphs

- Dataset size limitations
- No partition in results, require tree cutting

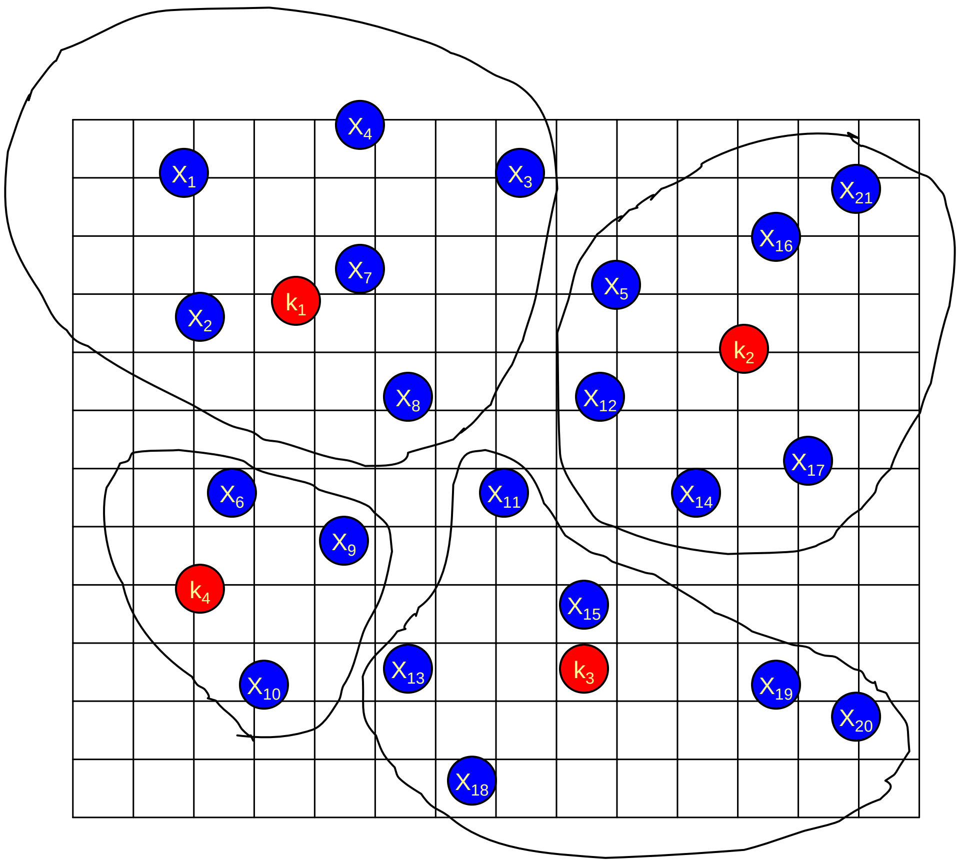
Partitional clustering : K-means

- Partition data into K clusters, with number K supplied by user.
- Produce cluster membership as results.

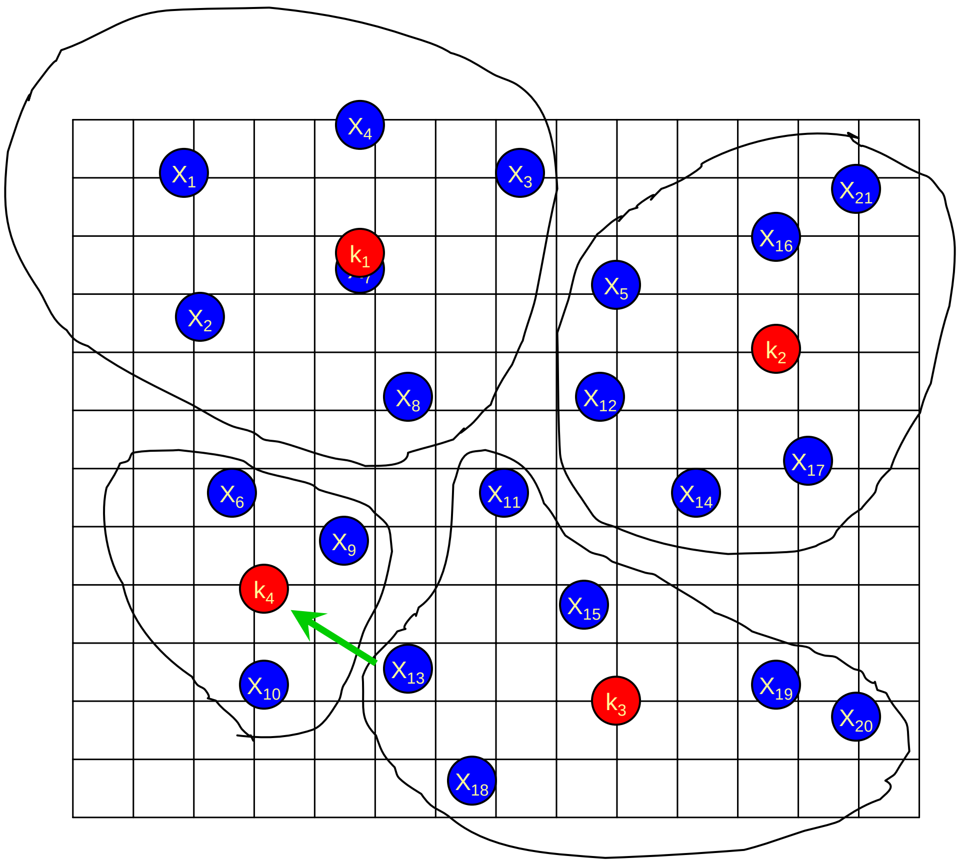
K-means Algorithm

- Divide observations into K clusters.
- Use cluster averages (means) to represent clusters
- Maximize the inter-cluster distance Minimize intra-cluster distance.

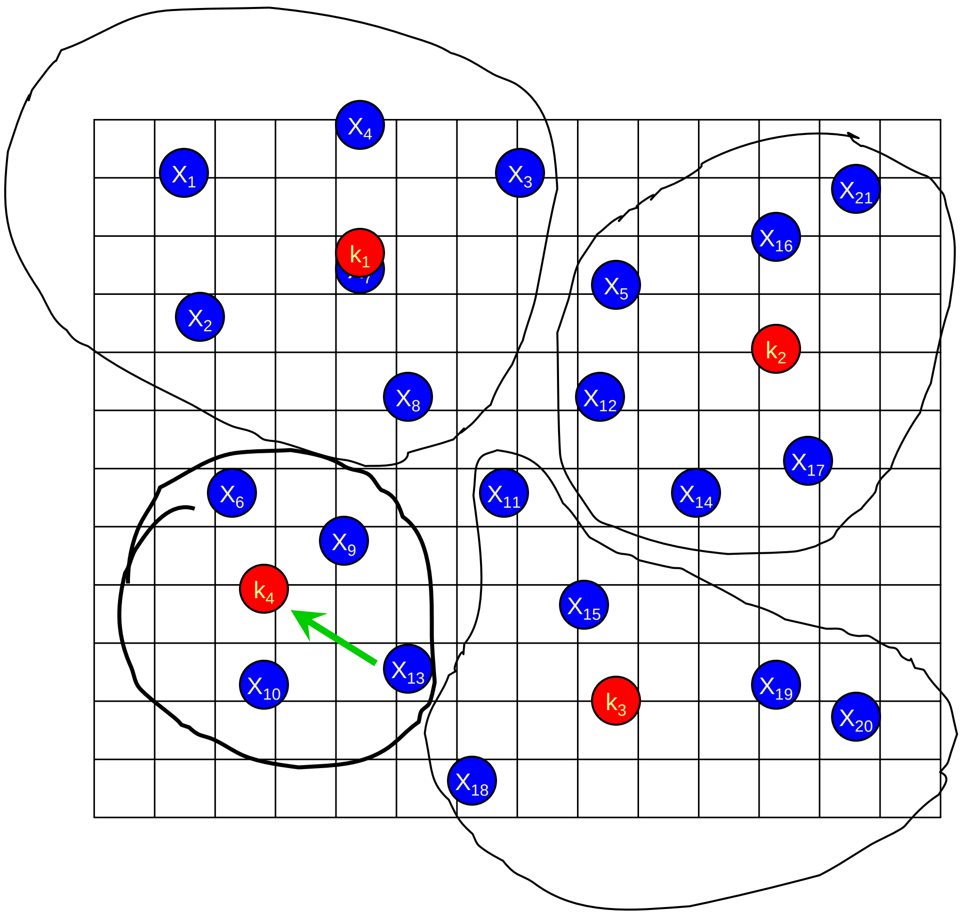
K-means Algorithm



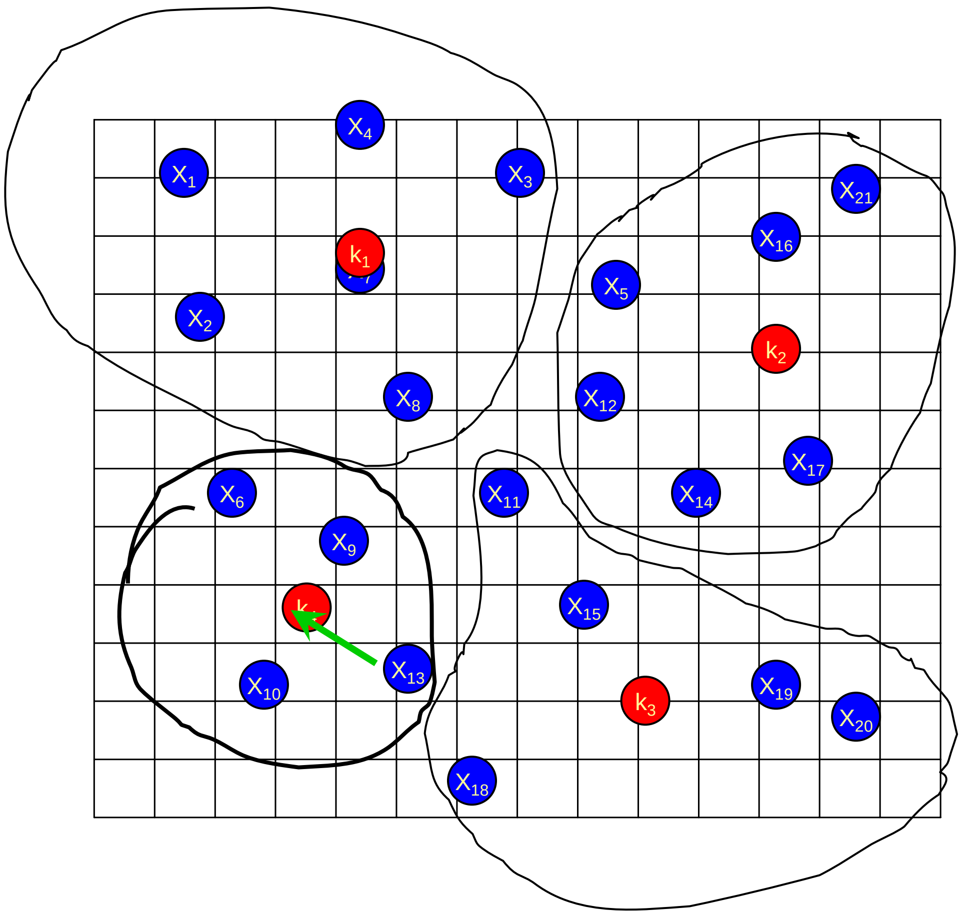
K-means Algorithm



K-means Algorithm



K-means Algorithm

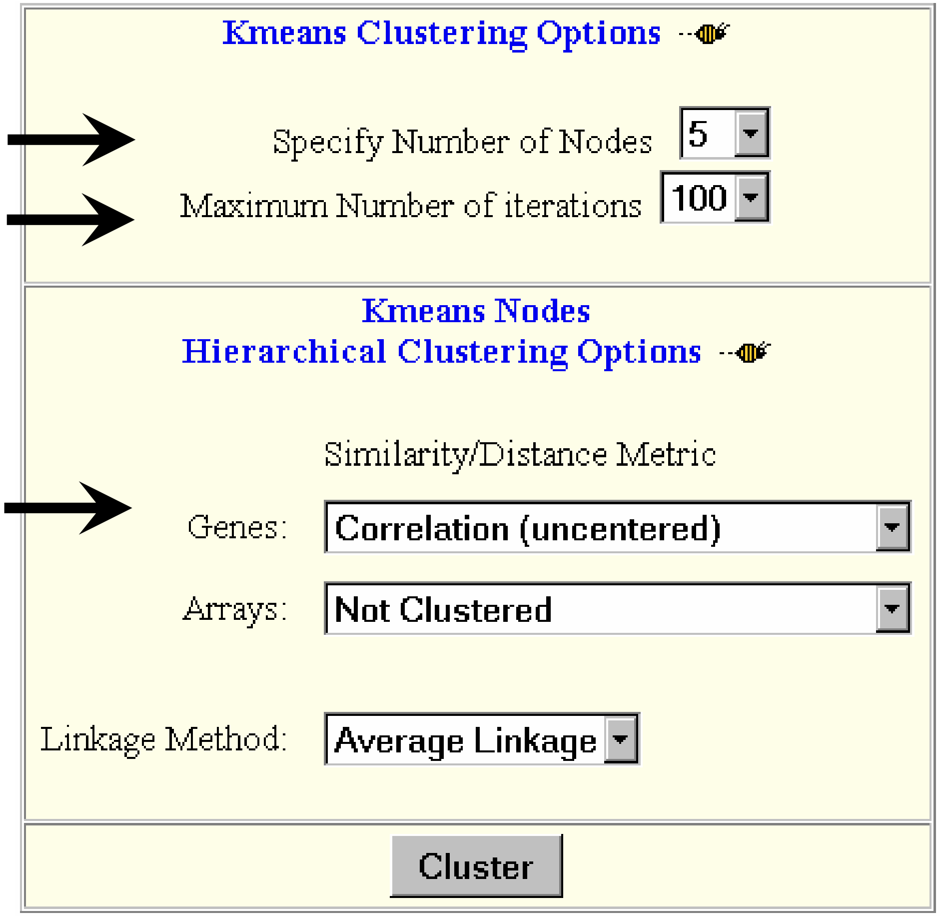


mAdb K-means Options

Set number of clusters

Set number of iteration

Hierarchical clustering within node

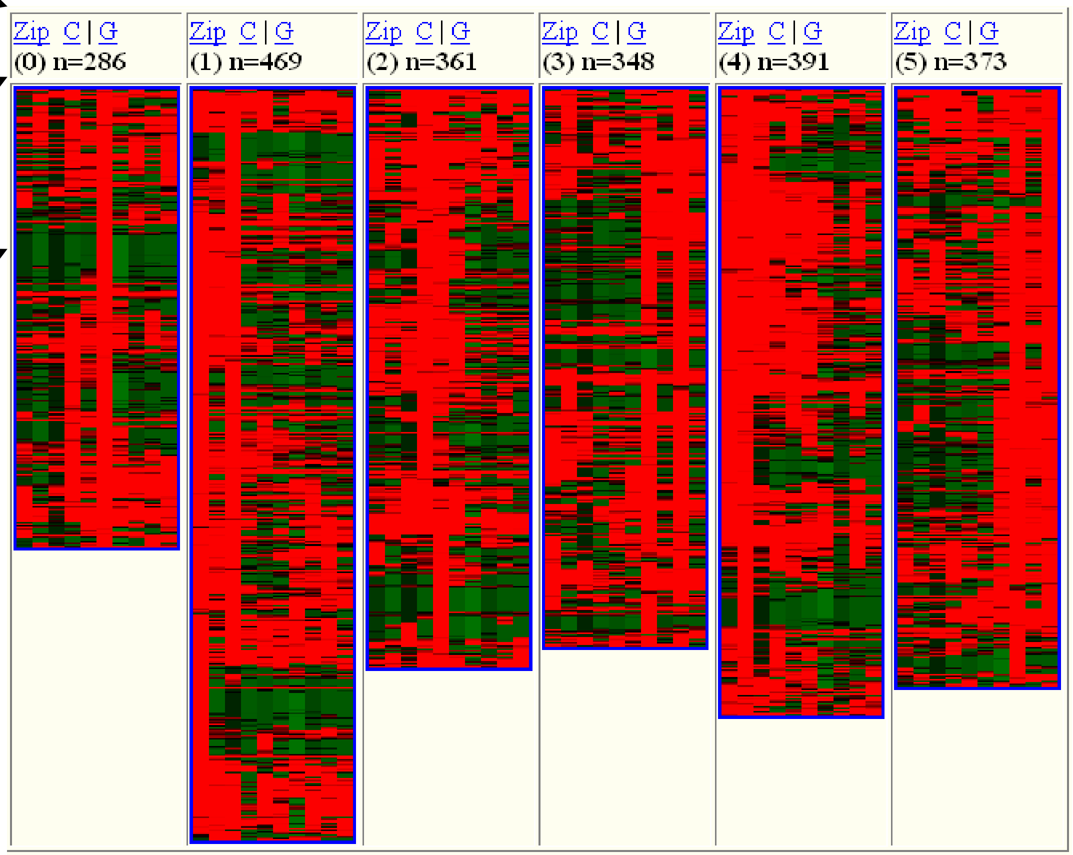


K-means Clustering Example

Save as input to TreeView

Create new subset of genes

Show hierarchical clustering



Summary

- Fast algorithm
- Partitions features into smaller, manageable groups
- mAdb allows hierarchical clustering within each K-mean cluster

- Must supply reasonable number of K
- No relationship among partitions

Self-Organizing Maps (SOM)

- Partitions data into 2 dimensional grid of nodes
- Clusters on the grid have topological relationships

- 2 numbers for the dimension of grid supplied by user

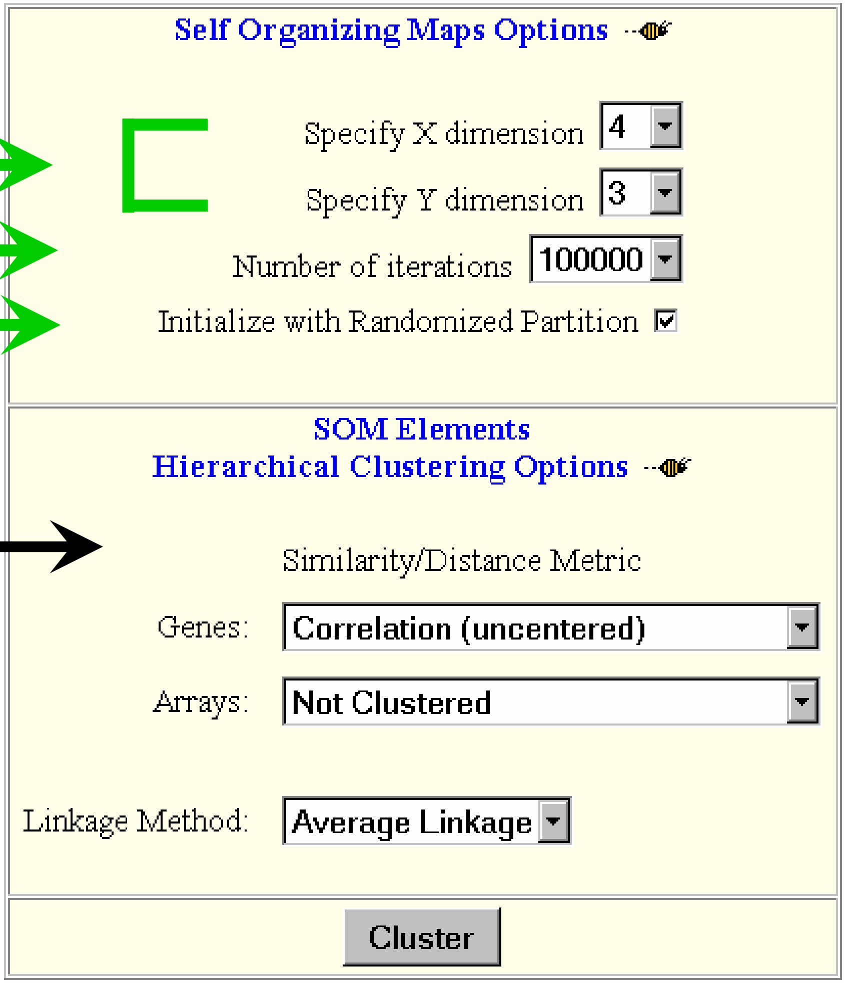
mAdb SOM options

Set number of clusters (X, Y)

Set number of iteration

Activate Randomized Partition

Hierarchical within SOM clusters

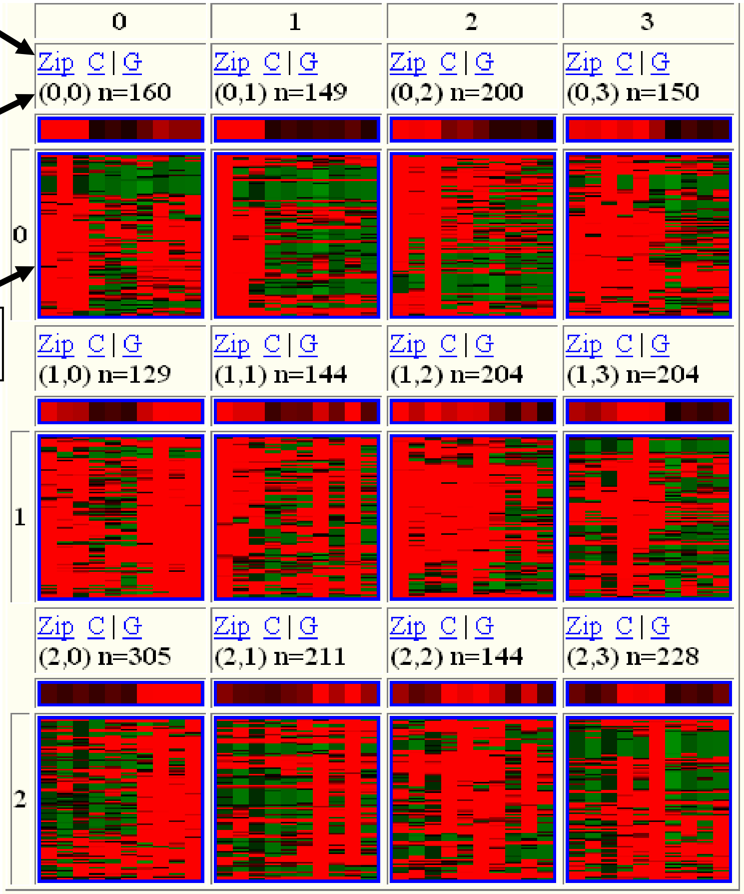


SOM Clustering Example

Save as input to TreeView

Create new subset of genes

Show hierarchical clustering



SOM Summary

- Neighboring partitions similar to each other
- Partitions features into smaller groups
- mAdb allows hierarchical clustering within each SOM cluster

- Results may depend on initial partitions

Summary of mAdb Clustering Tools

	Hierarchical	K-means	SOM
Relationship visualization	Tree Structure	partition Membership	Partition 2-D topology
Data Size	Small	Large	Large
Performance	Slow	Fast	Middle
Cluster Type	Gene/Array	Gene	Gene

Cluster Analysis

- Normalization is important
- Reduce data points by variance
- Use K-mean or SOM to partition dataset
- Use biological information to interpret results

Hands-on Session 2

- Lab 6 - lab 7
- Total time: 15 minutes

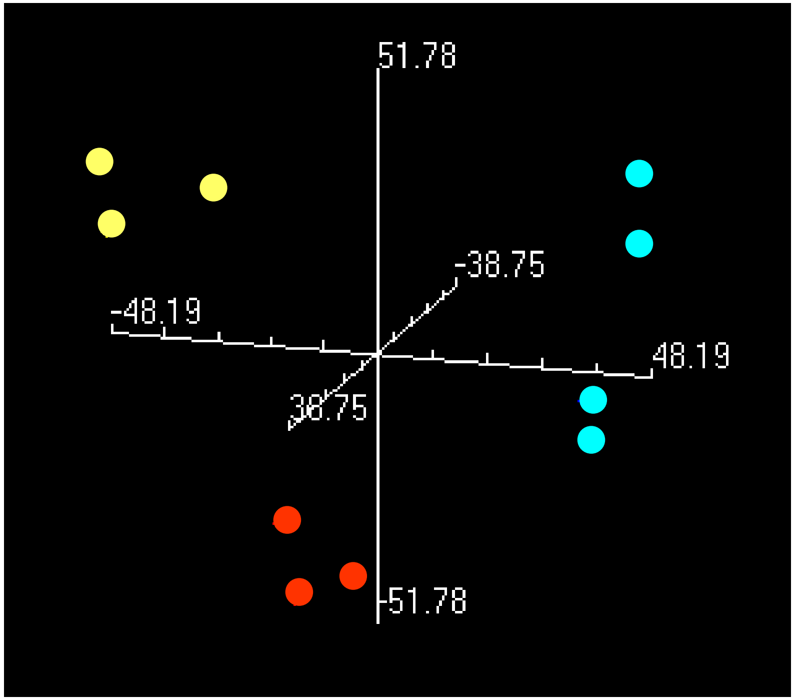
Principal Component Analysis

- How different samples are from each other
- Project high-dimensional data into lower dimensions, which captures most of the variance
- Display data in 2D or 3D plot to reveal the data pattern

Principal Component Analysis

- Hypothesis - there exist unobservable or "*hidden*" variables (complex traits) which have given rise to the ***correlation*** among the observed objects (genes or microarrays or patients)
- The Principal Components (PC) Model is a straightforward model that seeks to achieve this objective

PCA 3D plot



- Axes represent the first 3 components
- The first 3 components should explain most of the variance
- Formation of clusters
- Relationship of clusters.

Basic Idea of PCA is a Data Reduction Method Based on Analysis of Correlation Pattern(s) That Can Be Exist Among the Observed Random Variables (i.e. Expression values of Genes).

Raw Data

Array	1	2	…	m
Gene 1	a_{11}	a_{12}	…	a_{1m}
Gene 2	a_{21}	a_{22}	…	a_{2m}
Gene ...	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene n	a_{n1}	a_{n2}	…	a_{nm}

n is the number of genes (gene probes); m is the number of arrays (experiments)

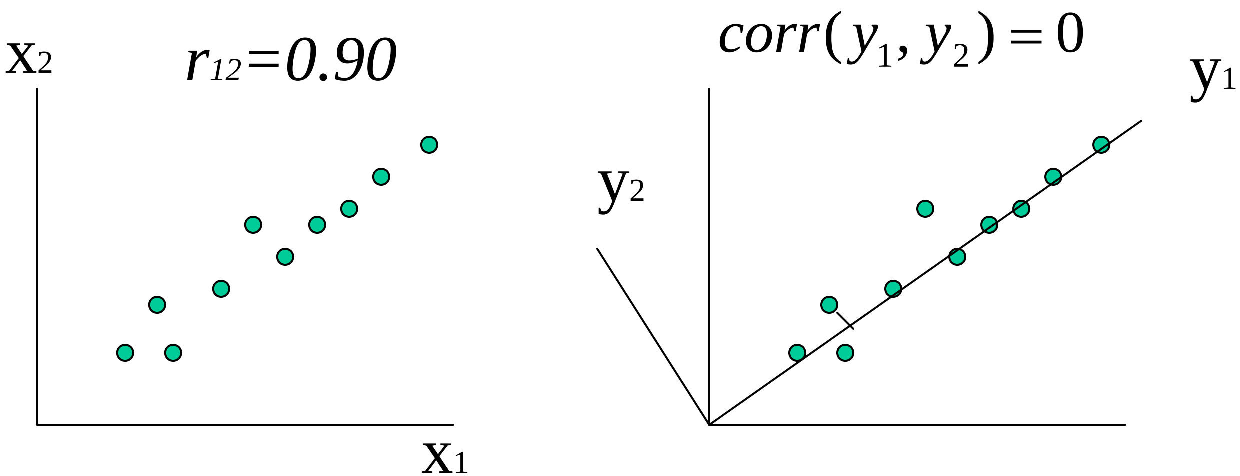
A Structure of Correlation Matrix is the **Major Object for PCA**

Correlation Matrix	Gene 1	Gene 2	…	Gene n
Gene 1	1	r_{12}	…	r_{1n}
Gene 2	r_{21}	1	…	r_{2n}
Gene ...	$\vdots$	$\vdots$	$\vdots$	$\vdots$
Gene n	r_{n1}	r_{n2}	…	1

A correlation matrix is a symmetric matrix of correlation coefficients
($-1 \le r_{ij} \le 1$ and $r_{ij} = r_{ji};\ i, j = 1,2,...,n;\ r_{ii} = 1$)

The Results of PCA are a small set of the orthogonal (independent) Variables Grouping of the Variables

From a purely mathematical viewpoint the purpose of PCA is to transform **n** correlated random variables to an orthogonal set which reproduces the original variance/covariance structure.

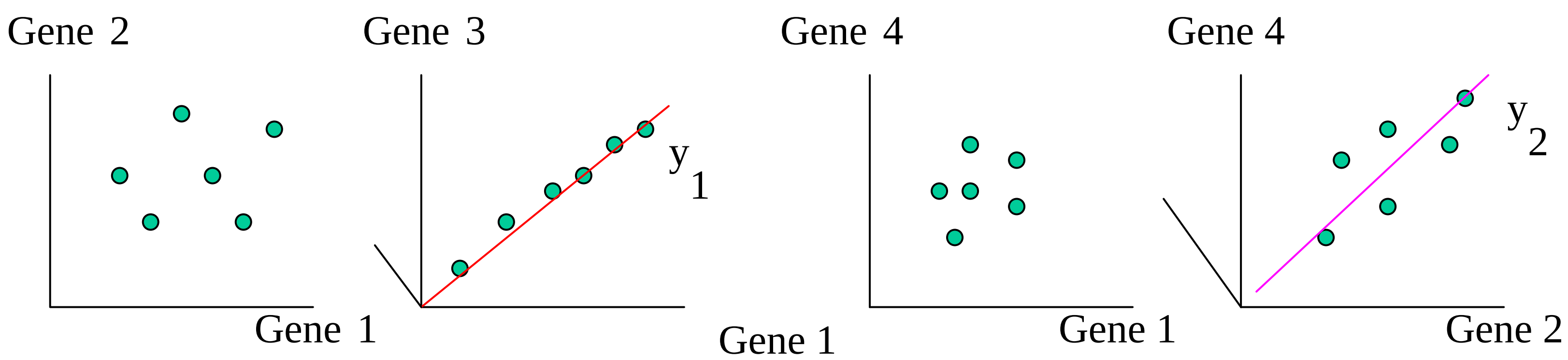


$$y_1 = b_{11}x_1 + b_{12}x_2$$
$$y_2 = b_{21}x_1 + b_{22}x_2 \approx 0$$

(The First) Principal Component y_1 can “explain” the major fraction (~90%) of a dispersion of variables x_1 and x_2 for all of the 10 observed objects.

The Example of a PC Model

Correlation Matrix	Gene 1	Gene 2	Gene 3	Gene 4
Gene 1	1	0.01	0.95	0.02
Gene 2	0.01	1	0.03	0.45
Gene 3	0.95	0.03	1	-0.03
Gene 4	0.02	0.45	-0.03	-0.03



$$y_1 = b_{11}x_1 + b_{13}x_3 + e_1(b'_{12}\, x_2 + b'_{14}\, x_4)$$

$$y_2 = b_{22}x_2 \;\; + b_{24}x_4 + e_2(b'_{21}\, x_1 + b'_{23}\, x_3)$$

$$y_3 = e_3(b'_{31}\, x_1 + b'_{32}\, x_2 + b'_{33}\, x_3 + b'_{34}\, x_4)$$

$$y_4 = e_4(b'_{41}\, x_1 + b'_{42}\, x_2 + b'_{43}\, x_3 + b'_{43}\, x_4)$$

$$corr(y_i, y_j) = 0; i \neq j$$

$$e_i << 1; i = 1,2,3,4 \longrightarrow$$

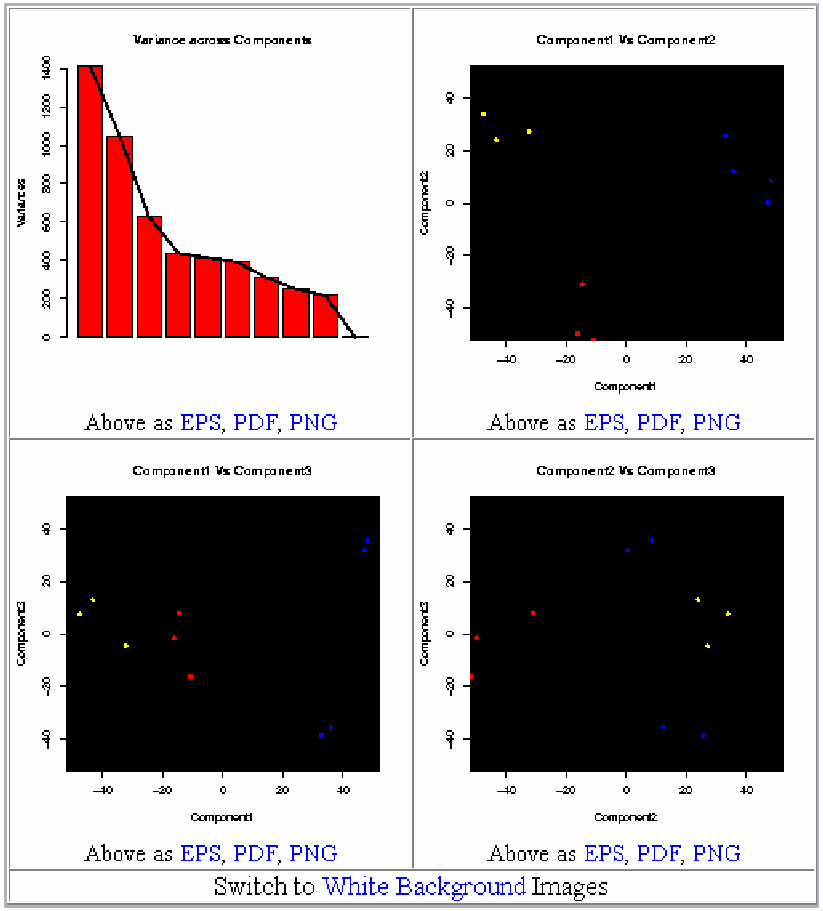
$$\boxed{y_1 \approx b_{11}x_1 + b_{13}x_3}$$

$$\boxed{y_2 \approx b_{22}x_2 + b_{24}x_4}$$

Sample:Small Round Blue Cell Tumors (SRBCTs)

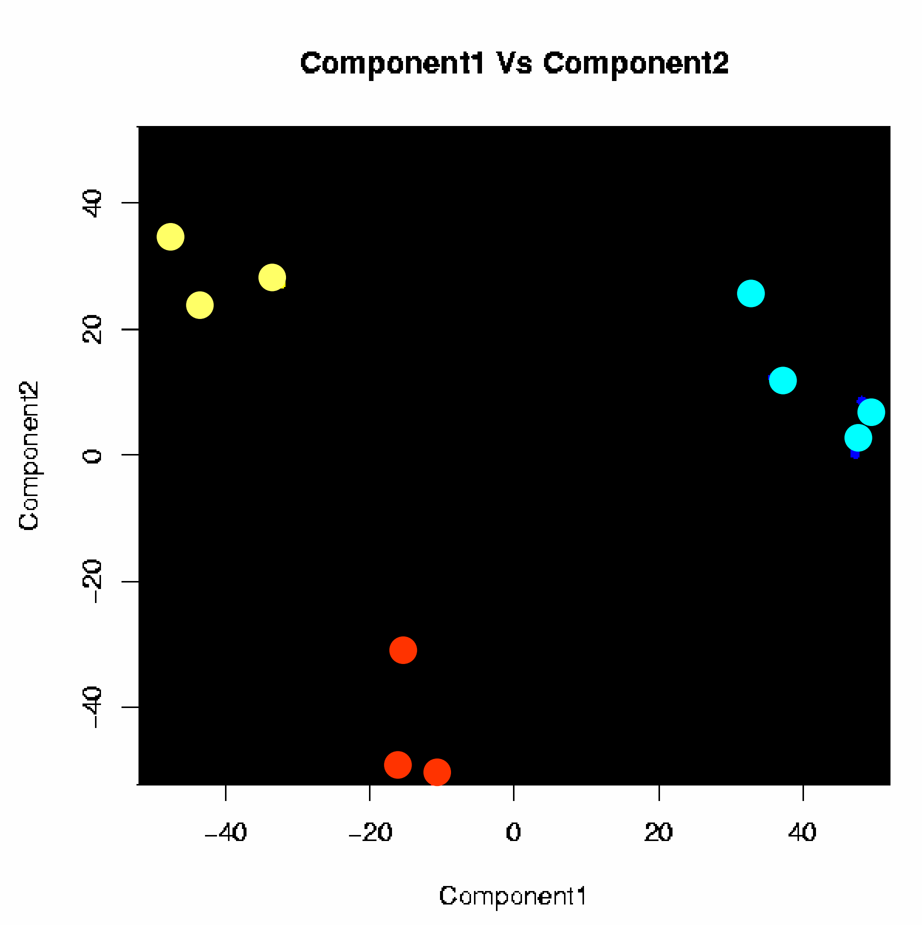
- 63 Arrays representing 4 groups
  - BL (Burkitt Lymphoma, n1=8)
  - EWS (Ewing, n2=23)
  - NB (neuroblastoma, n3=12)
  - RMS (rhabdomyosarcoma, n4=20)
- There are 2308 features (distinct gene probes)

PCA Detailed Plot



- ”Scree” plot
- 2-D plots

PCA 2-D plots



- First 2 components separate 3 groups well

Result of the PCA :

Comp1	Comp2	UniGene	Description
0.00934	0.000195	Hs.119571	collagen type III alpha 1=Ehlers-Danlos syndrome type IV autosomal dominant=COL3A1
0.008788	2.36E-05	Hs.78935	methionyl aminopeptidase 2
0.008736	5.49E-05	Hs.83164	collagen, type XV, alpha 1
0.008063	5.30E-05	Hs.180324	Human insulin-like growth factor binding protein 5 (IGFBP5) mRNA
0.007908	0.000521	Hs.251664	Homo sapiens cDNA: FLJ22066 fis, clone HEP10611
0.007408	0.000288	Hs.349109	Insulin-like growth factor 2 (somatomedin A)
0.006517	0.000498	Hs.78846	heat shock 27kDa protein 2
0.005894	0.000107	Hs.374415	ESTs
0.005651	9.83E-06	Hs.290070	gelsolin (amyloidosis, Finnish type)
0.005402	0.0001	Hs.15463	Homo sapiens, clone IMAGE:2959994, mRNA
0.005047	0.000121	Hs.84520	Yes-associated protein 1, 65kDa
0.005012	0.000389	Hs.151242	serine (or cysteine) proteinase inhibitor, clade G (C1 inhibitor), member 1, (angioedema, hereditary

Comp1	Comp2	UniGene	Description
3.96E-05	0.01071	Hs.73853	bone morphogenetic protein 2
6.47E-05	0.010634	Hs.89709	glutamate-cysteine ligase, modifier subunit
4.63E-05	0.008607	Hs.239760	citrate synthase
9.14E-05	0.008508	Hs.31053	cytoskeleton-associated protein 1
0.000428	0.008408	Hs.174195	interferon induced transmembrane protein 2 (1-8D)
0.00038	0.008193	Hs.159637	valyl-tRNA synthetase 2
8.30E-05	0.007452	Hs.79876	steroid sulfatase (microsomal), arylsulfatase C, isozyme S
1.20E-05	0.007	Hs.43509	ataxin 2 related protein
0.003848	0.006756	Hs.303627	heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)
1.30E-05	0.00652	Hs.106876	ATPase, H+ transporting, lysosomal 38kDa, V0 subunit d isoform 1
7.68E-06	0.006387	Hs.290791	ESTs
0.002802	0.006052	Hs.289271	cytochrome c-1

MDS overview

- An alternative for PCA
- Non-linear projection methodology
- Tolerates missing values

Summary of PCA and MDS

- Dimension reduction tools
- Graphic representation to help explain patterns
- Quality control for experimental variance

Hands-on Session 3

- Lab 5
- Total time: 15 minutes

- Next class tomorrow at 1:00 pm

Agenda

1. mAdb system overview
2. mAdb dataset overview
3. mAdb analysis tools for dataset
   - Class Discovery – clustering, PCA, MDS
   - Class Comparison-statistical analysis
   - Class Prediction –PAM

Various Hands-on exercises

Class Comparison

- Overview – Statistical distributions and statistical tests
- Statistical Distributions of Gene expression and Microarray Data Analysis
- Hypothesis tests for two or more groups
  - Errors: Type 1 and Type 2
- mAdb analysis tools – Statistical tests
  - T-test
  - ANOVA

Sources of errors and uncertainty in microarray data analysis

- Poorly-controlled external factors (quality of tissue sample, RNA etc.)
- Mixture of biological samples derived from many cells and/or complex tissues
- Biological noise (stochastic mechanisms of gene expression)
- Technical Noise of background signals
- Inter-array and across- array normalizations.
- Limited number of replicates (cost, personnel, etc. constraints)
- Inadequate statistical methods

Class Comparison

Goal: To introduce users to some basic statistical tests and data mining tools in mAdb to identify differentially expressed genes

Gene Expression Levels

The **gene's expression level** is defined as the ***average number*** of mRNA molecules per cell.

A complete list of mRNAs of a given cell type
is called the ***transcriptome*** **.** Observed list of mRNAs in the RNA sample is called the ***representative transcriptome of a cell population.***

Differentially Expressed Genes

The goal of testing for differentially expressed genes is the identifying a complete list of genes having expression levels statistically and (more important) biologically different in two or more sets of the representative transcriptomes.

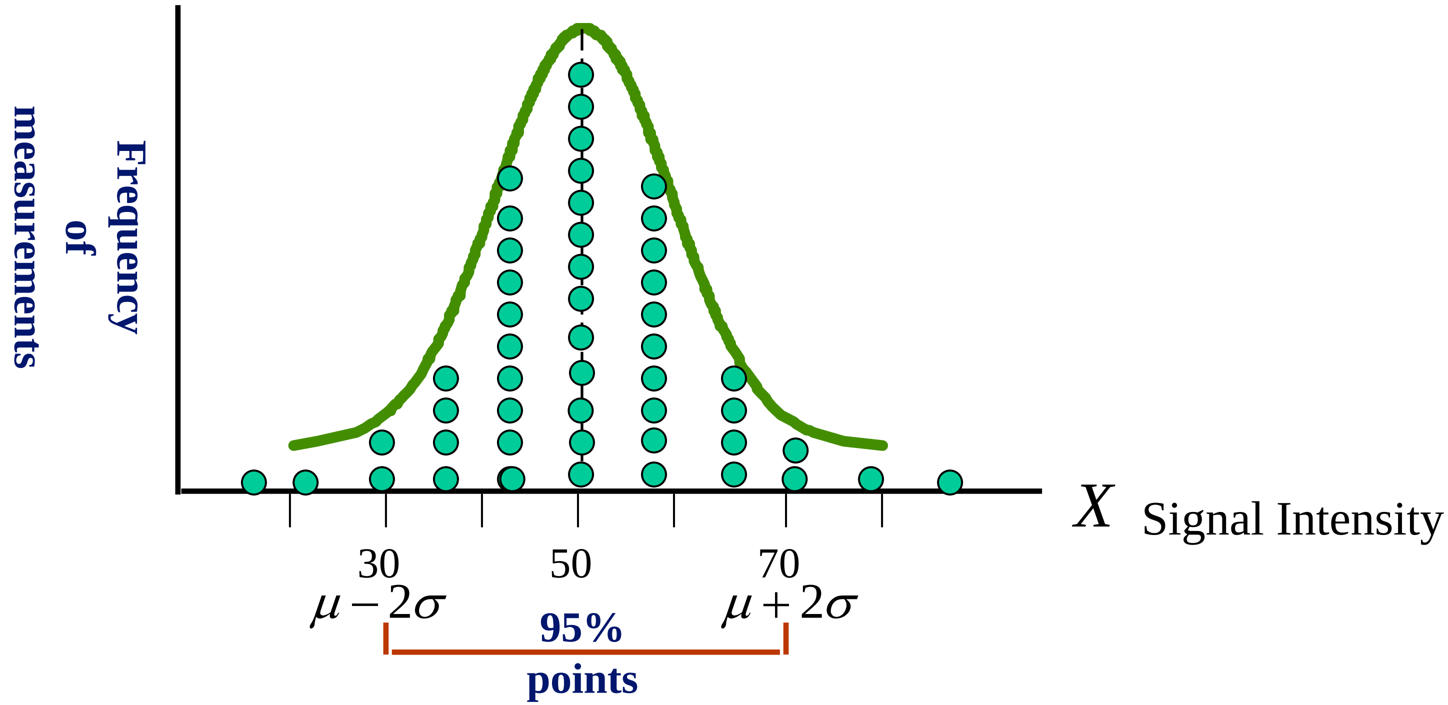
A frequency concept of probability

Let **n(A)** be the number of occurrences of event A in the **N** repetitions of the same experiment.
The frequency concept states that the ratio n(A)/N approximates the probability P(A) of even A with accuracy of the approximation increasing as N increases.Thus, the probability of an event A is additively countable, non-negative value in the closed interval [0,1].

An estimate of P(A)= (number of occurrences of A)/Total number of occurrences

To present the probabilities of all possible events of the experiment, we can construct the histogram (the empirical frequency distribution)
which approximates the **probability function** of a random variable associated with these events.

Replicated measurements and the Frequency Distribution Function

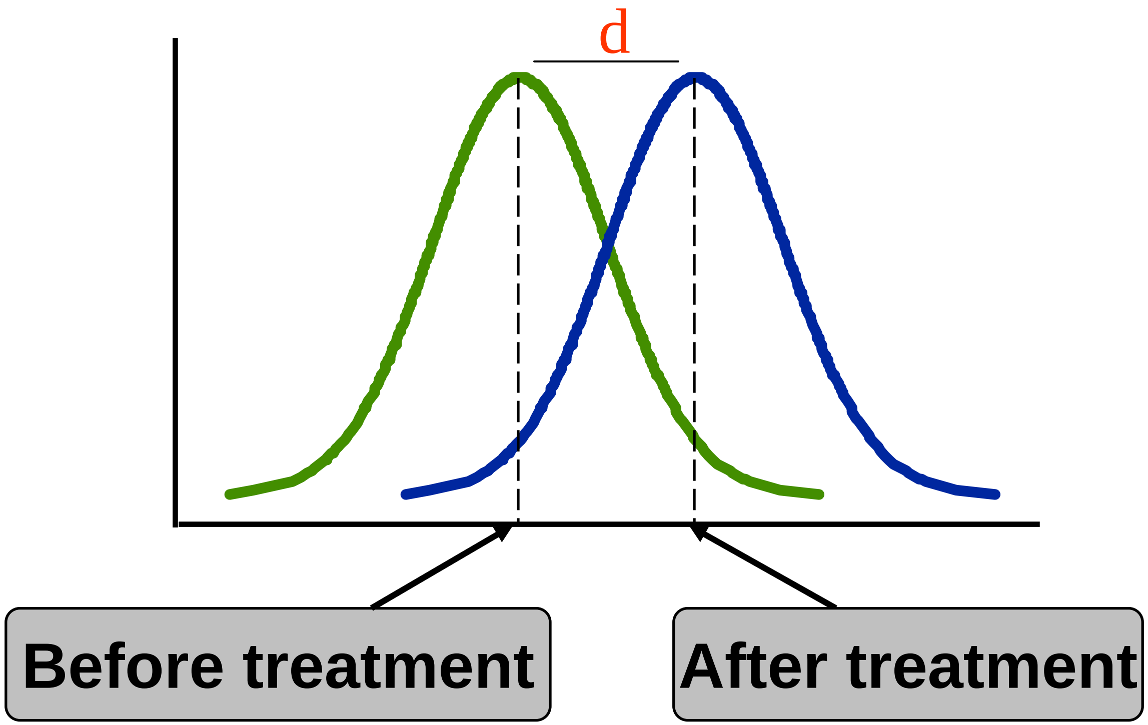


Sampling of Normal Distribution: $f(x) = P(X = x) = \frac{1}{\sigma(2\pi)^{1/2}} \exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)$

Mean $\mu = (1/N)\sum_{i=1}^{N} x_i$ Standard deviation $\sigma = [(1/(N-1))\sum_{i=1}^{N}(x_i - \mu)^2]^{1/2}$

N=Number of observations (sample size)

Testing the hypothetical frequency distributions of the expression level for a gene in two populations



Null hypothesis $H_o : \mu_1 = \mu_2 ; \sigma_1 = \sigma_2$

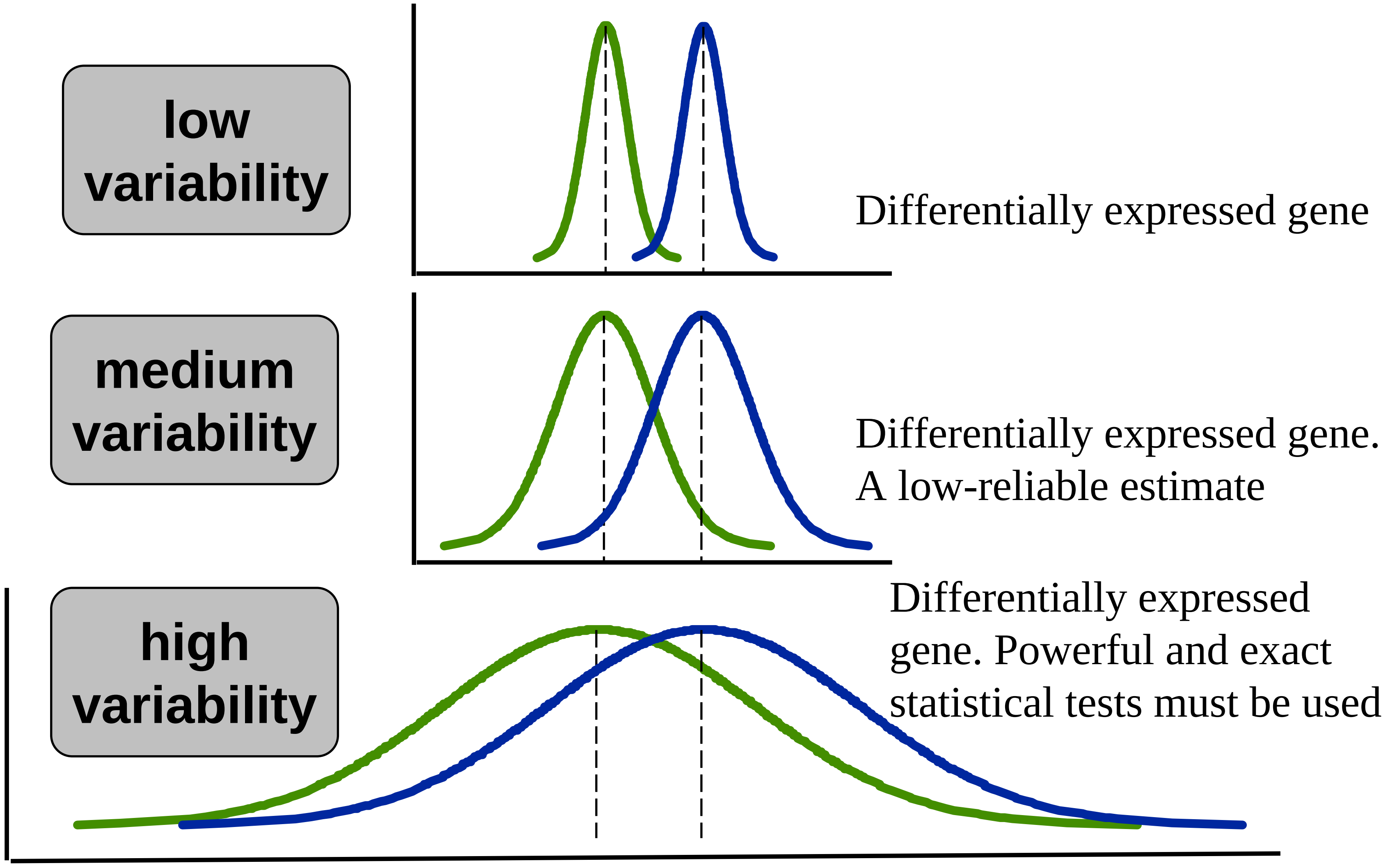
Alternative hypotheses

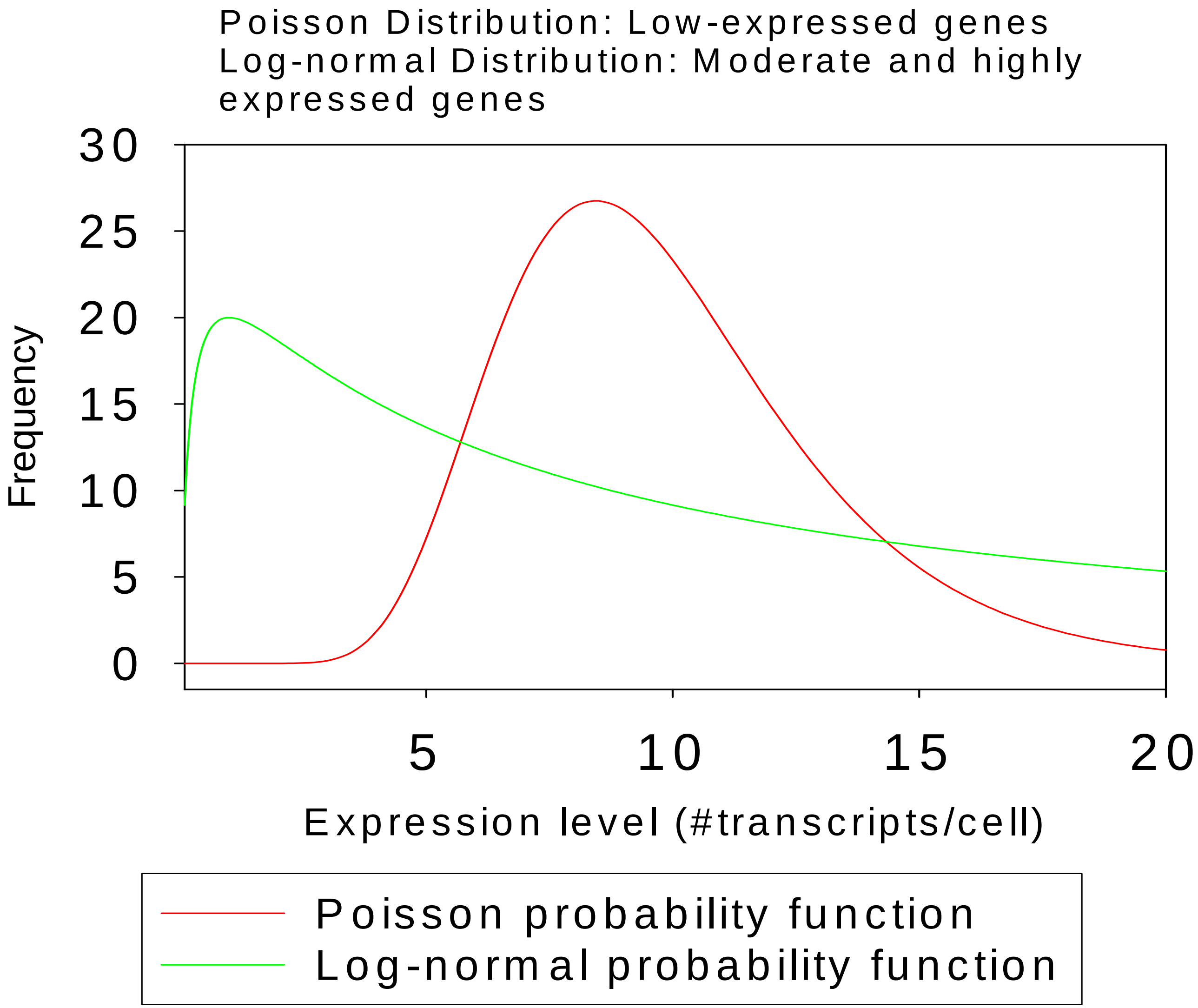
$$H_1 : \mu_1 \neq \mu_2 ; \sigma_1 = \sigma_2 ;$$

$$\mu_1 = \mu_2 ; \sigma_1 \neq \sigma_2 ;$$

$$\mu_1 \neq \mu_2 ; \sigma_1 \neq \sigma_2 ;$$

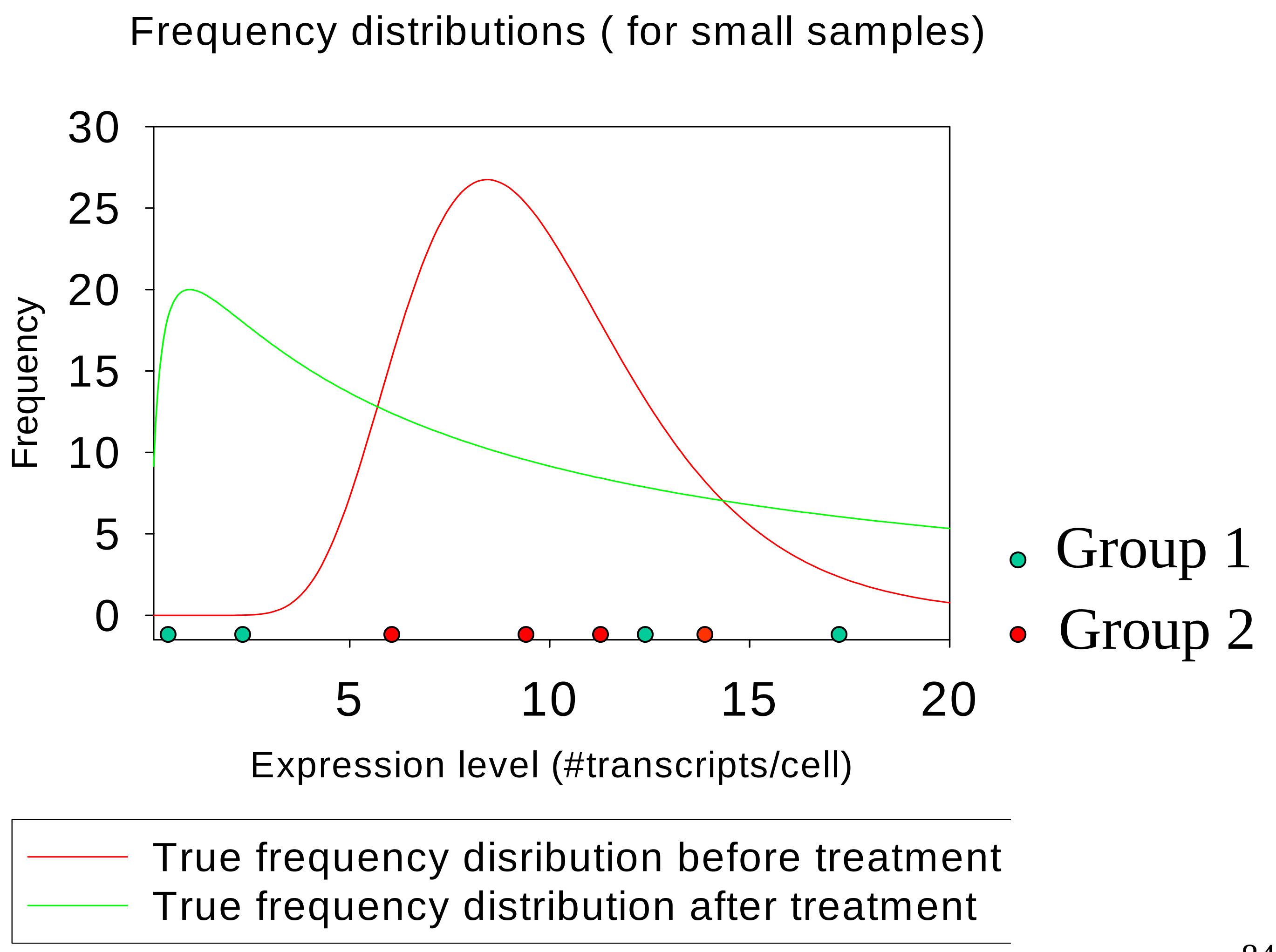
Spread (variability) of measurements





Poisson Distribution: Low-expressed genes
Log-normal Distribution: Moderate and highly expressed genes
Frequency
30
25
20
15
10
5
0
5
10
15
20
Expression level (#transcripts/cell)
Poisson probability function
Log-normal probability function

Comparison of the Skewed Distributions: A Problem of Sample Size



Gene Expression Profile

• Gene expression data sets have very broad ranges of the number of transcripts for different genes (from 0.1 to 20000 transcripts per human cell on average).

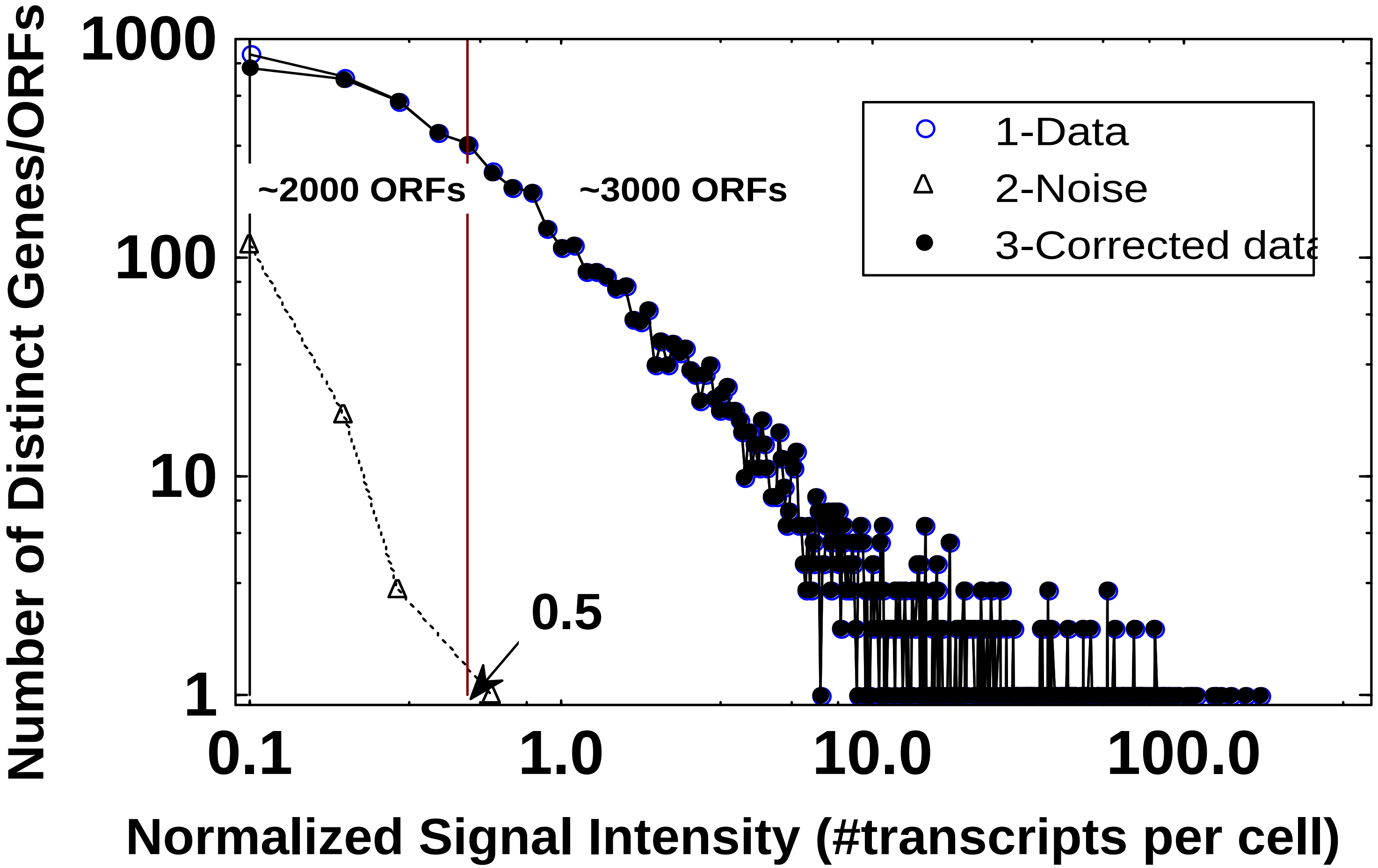
• The list of the mRNA transcripts found in the representative transcriptome, together with each gene’s expression level is called the ***gene expression profile.***

Statistical Distribution of Gene Expression Levels

• The ***statistics of expressed genes*** can be specified by the number (and/or proportion) of expressed genes that have one, two, etc. transcripts present in an associated mRNA sample.

• A normalized histogram of gene expression levels can be considered as the empirical frequency distribution of the numbers of expressed genes

Typical skewed frequency distribution of the gene expression levels in the eukaryotic transcriptome (Kuznetsov, VA. et al., Genetics, 161, 1321-1332, 2002)



Determination of the working domain for signal intensity levels in which differentially expressed genes might be found. By our estimates ~ 40% (2000 genes) of the 5000 apparently expressed yeast genes are expressed at less than 0.5 copy per cell on average.

Frequency Distribution of Gene Expression: Observations

A frequency distribution of the gene expression levels in the transcriptomes has skewed shape with a very long right tail.

Statistical analysis implies that most of the expressed genes in eukaryotic cells have few transcripts per cell

Technical Caveats

• Technical variability (noise) has a significant intensity bias toward low signal intensity values

• Simple, static fold change thresholds are too stringent at high intensities and not stringent enough at low intensities.

Statistical and biological problems with fold change of means:

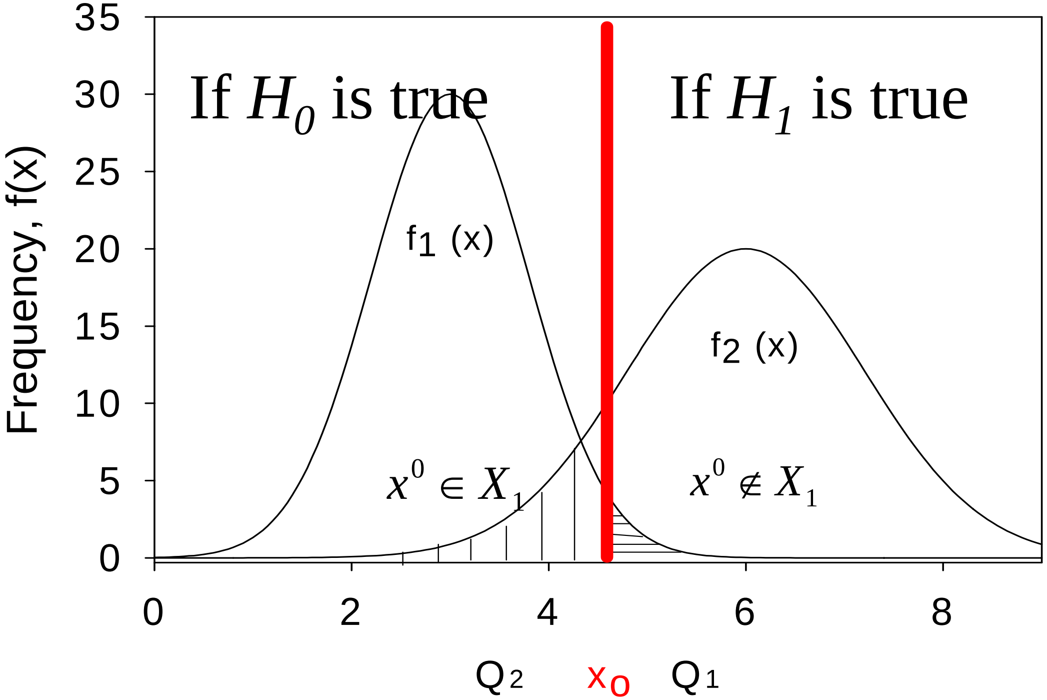
- Genes with high fold change may exhibit high variability among cell types due to natural biological variability for these genes
- Genes with small fold changes may be highly reproducible and should be biologically essential genes

Conclusion:

Robust Statistical Tests of microarray data are necessary to use and an additional Biological validation(s) of the statistical analysis should be needed

Hypothesis tests for two or more groups

Two types of Errors



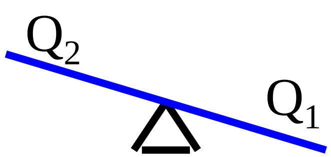
X_1= data set for control population;
X_2 = data sets for tested population.
Let x_0 **be the critical (the rejection) value of** x. Let x^o be the observed value of x.

If x^o belongs X_1, then ***deciding*** that x^o not belongs X_1 is the **error of type I**.

If x^o belongs X_2, then ***deciding*** that x^o not belongs X_2 is the **error of type II**.

H_0: Q_1=The probability of an error of type I (false-positive)

H_1: Q_2=The probability of an error of type II (false-negative)



False-negative *False-positive*

Any modifications of x_0 has the opposite effects on probabilities of errors of Type I and Type II: if Q_1 is pushed down, then Q_2 is raised. However, an increase of sample size decreases of both types of errors.

The Decision:
Relation Between Type I and Type II Errors

	Accept *Ho*	Reject *Ho*
Ho is true	Correct decision Probability=1-Q_1	Type 1 error Probability=Q_1
Ho is false	Type II error Probability= Q_2	Correct decision Probability=1-Q_2 (power)

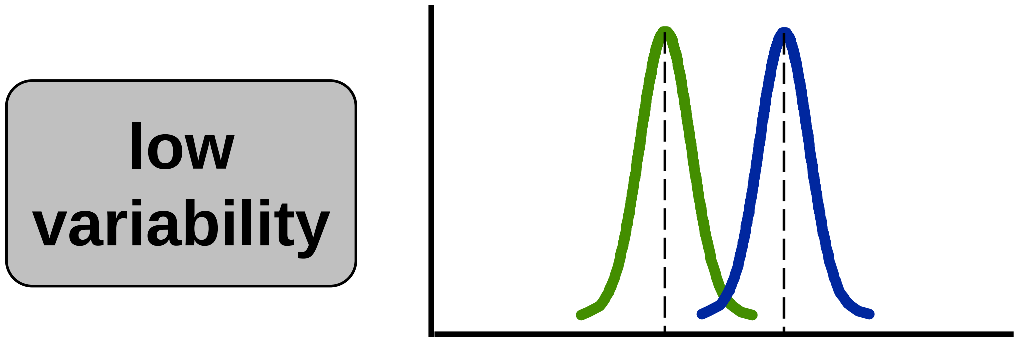
The ***p-value*** is the smallest probability (significance value) at which the *Null Hypothesis* , *Ho*, would be rejected by a test for a given data

The t-test assesses whether the means of two groups are statistically different

The null hypothesis is

$$H_o : \mu_1 - \mu_2 = 0$$

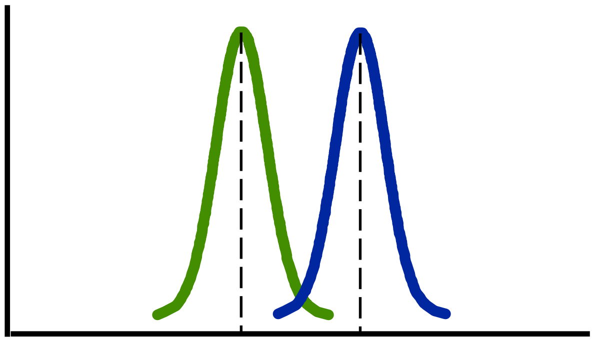
Calculating t-test



Compare the means of two groups

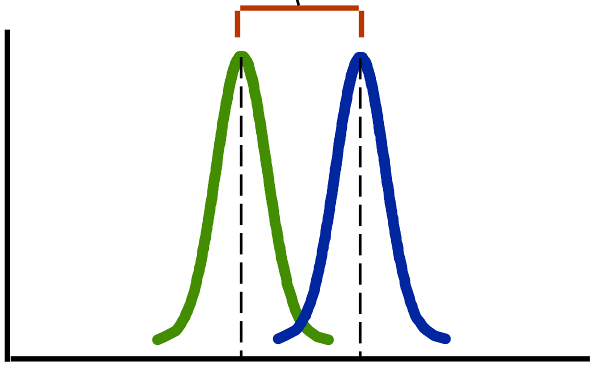
$$\frac{\text{signal}}{\text{noise}}$$

low variability



Compare the means of two groups

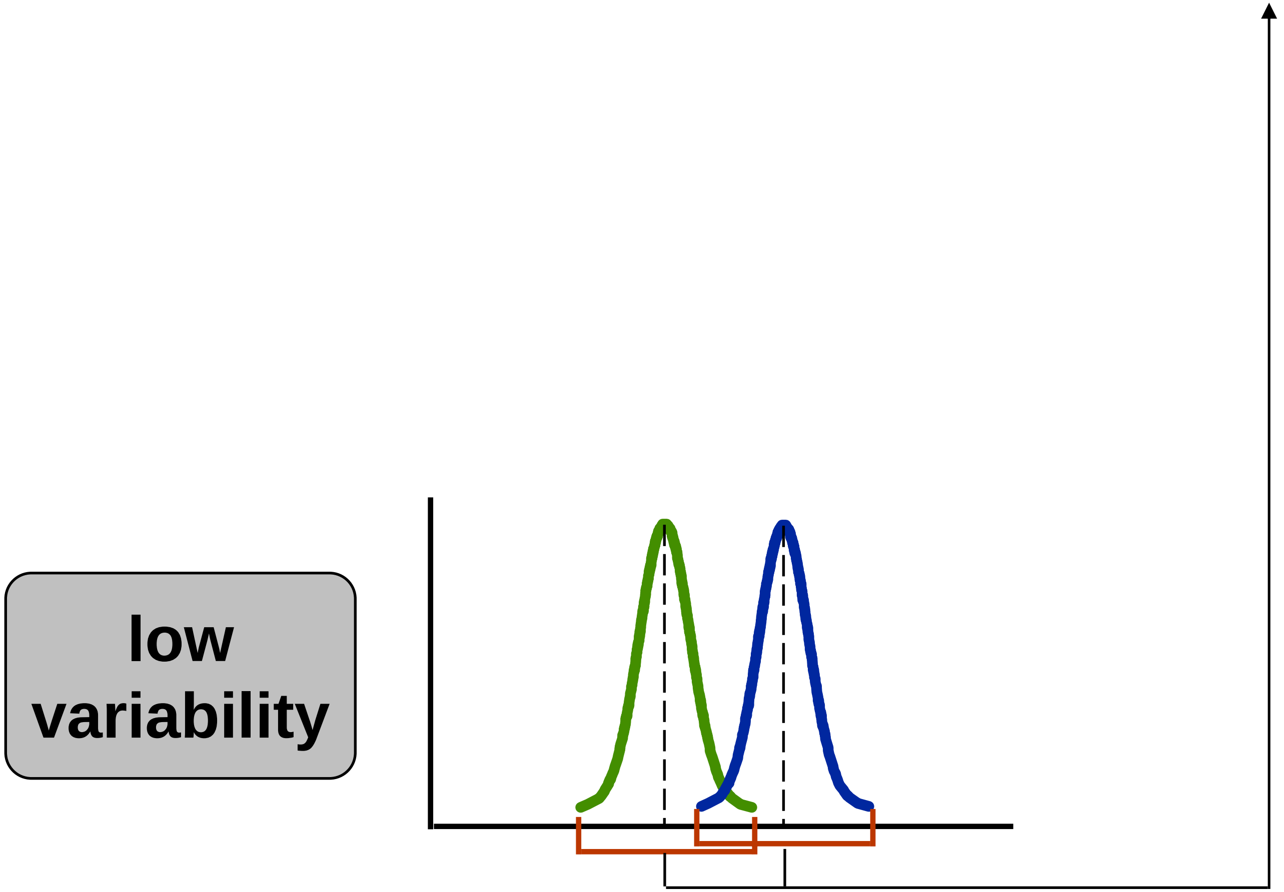
$$\frac{\text{signal}}{\text{noise}} = \text{difference between group means}$$



low variability

Compare the means of two groups

$$\frac{\textbf{signal}}{\textbf{noise}} = \frac{\textbf{difference between group means}}{\textbf{variability of groups}}$$

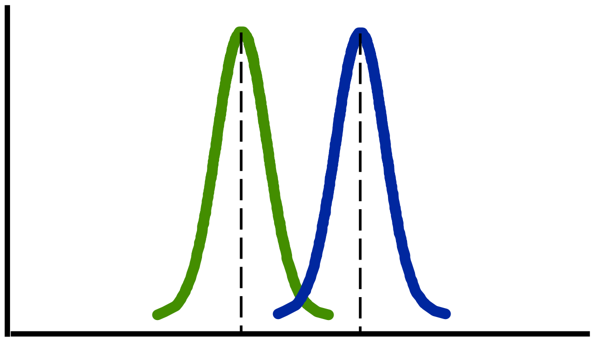


Compare the means of two groups

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

low variability



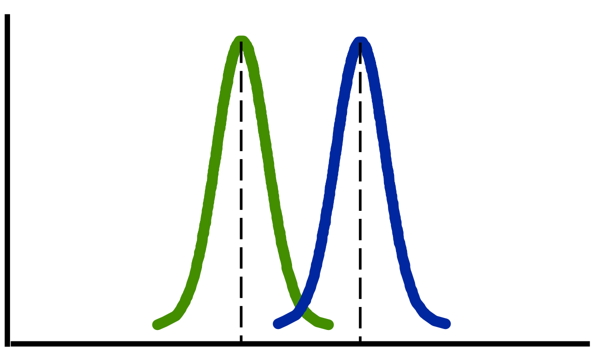
Compare the means of two groups

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}}$$

$$= \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)}$$

$$= \text{t-value}$$

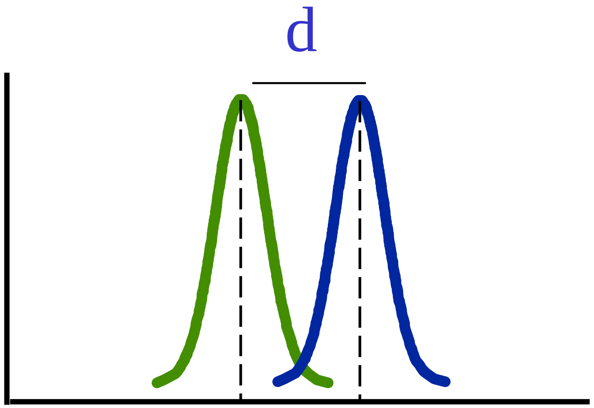
low variability



Calculating p-value (t-test)

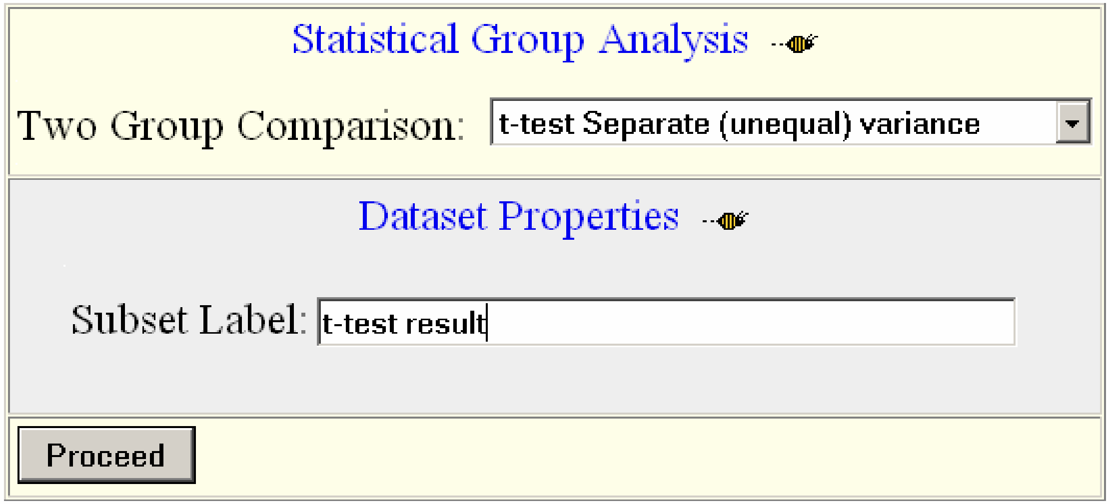
- The p-value is the probability to reject the null hypothesis ($H_o : \mu_1 - \mu_2 = 0$) when it is true (e.g. p=0.0001)
- When carrying out a t-test, a p-value can be calculated based on the *t*-value and the sample sizes n_1 and n_2.

Large distance d, low variability,



mAdb t-test

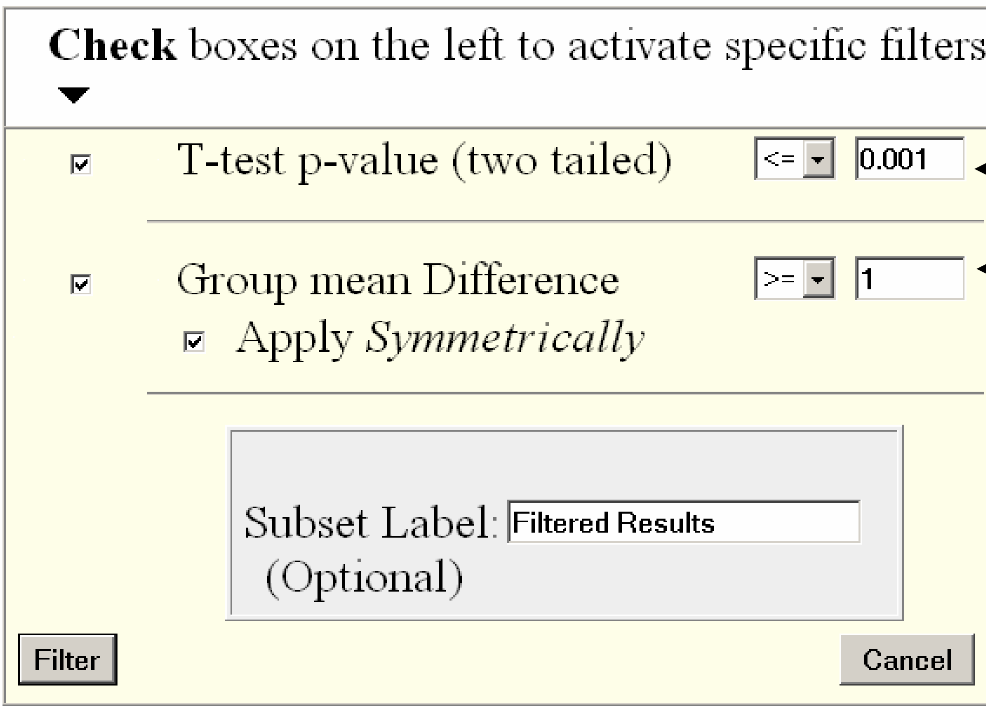
- 2 group statistic analysis automatically selected for a 2 group dataset



t-test Results

A	A	A	B	B	B		
JIM3_A	JJN3_A	U266_A	HDLM2_A	L428_A	L540_A	p-Value	Difference
52.4309	54.9520	45.0046	0.7800	0.6485	0.8532	1.9737e-06	6.07
35.1142	52.4541	42.8235	0.7800	0.6485	0.8532	8.9006e-06	5.83
53.3166	74.5535	46.5118	0.7800	0.6485	0.8532	1.1662e-05	6.24
5.9693	5.9444	5.7954	9.4782	9.6511	10.0555	1.4619e-05	-0.72
12.2739	13.0063	9.6026	0.7800	0.6485	0.8532	2.4704e-05	3.93
0.6680	0.6954	0.6536	9.0445	8.4780	13.0657	3.7853e-05	-3.9
3.7943	3.4277	3.3739	7.3190	7.6012	7.2551	4.7738e-05	-1.07
0.6680	0.6954	0.6536	2.3401	2.0402	2.5358	4.9127e-05	-1.77
0.6680	0.6954	0.6536	7.6466	6.0506	9.6493	5.7477e-05	-3.51
0.6680	0.6954	0.9490	8.0788	8.5636	6.8106	5.8369e-05	-3.35
0.6680	0.6954	0.7869	68.9017	34.0804	72.9403	6.3509e-05	-6.28
34.7315	29.5014	60.8882	0.7800	0.6485	0.8532	7.1258e-05	5.71
0.6680	0.6954	0.6706	0.8424	0.8593	0.8532	8.4299e-05	-0.329
0.6680	0.6954	0.6536	39.1841	17.6407	27.2176	9.1539e-05	-5.31
3.7288	2.9875	3.1098	0.9774	0.8392	0.8532	9.9425e-05	1.88
0.6680	1.3275	0.6536	26.2949	22.3119	26.9078	0.00014347	-4.91
1.7328	1.8435	2.0412	0.8557	0.9196	0.8532	0.00014599	1.09

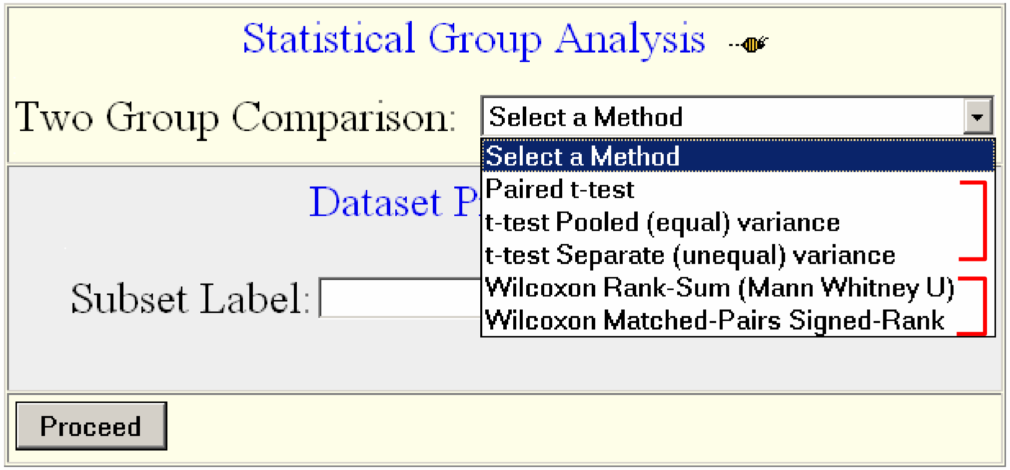
Statistic Results Filtering



statistical significance, p-value

$\mathbf{\log_2(x_1) - \log_2(x_2)}$

Other Statistical Tests for Univariate Analysis



Parametric

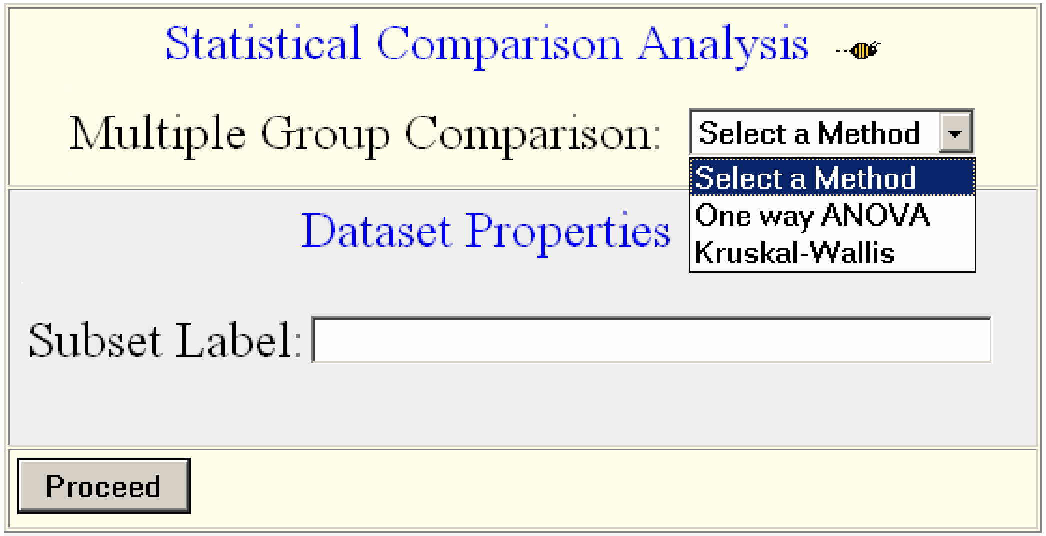
Non-Parametric (distribution free)

Analysis of the k independent groups (k>=2)

Group 1	Group 2	…	Group k
$X_{1,1}$	$X_{2,1}$	…	$X_{k,1}$
$X_{1,2}$	$X_{2,2}$	…	$X_{k,2}$
…	…	…	…
$X_{1,n1}$	$X_{2,n2}$	…	$X_{k,nk}$

H_o: All of the populations are identical;
H_1: Some of populations tend to display differ observed values than other populations

Multiple Group Comparison for each gene



- Analysis Of Variance (ANOVA): parametric test based on F-statistics
- Kruskal-Wallis : non-parametric rank-based test

Analysis of Variances (ANOVA)

This parametric method can be applied to compare several population means

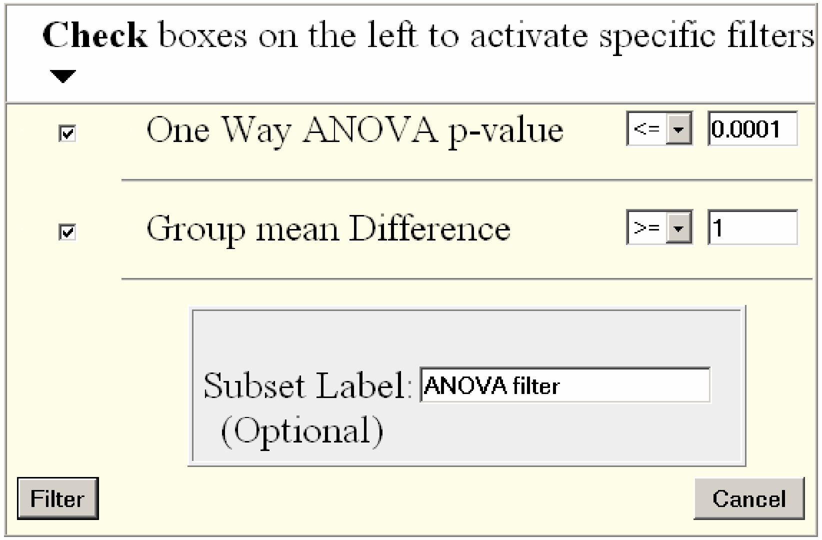
$$H_o : \mu_1 = \mu_2 = ... = \mu_k$$

vs.

$$H_1 : \mu_i \neq \mu_j; \quad \text{for some} \quad 1 \leq i \neq j \leq k$$

ANOVA Results and Filtering

p-Value	Difference	Groups
9.6276e-22	4.11	A-B
3.488e-20	2.99	D-C
2.5008e-19	3.59	A-B
2.5733e-18	2.59	A-D
1.4459e-17	2.76	D-A
5.7703e-17	2.89	A-B
8.728e-17	3.14	D-B
1.3957e-16	3.95	C-A
4.1114e-16	4.03	A-B
1.4464e-15	3.76	A-B
2.369e-15	3.1	D-B
7.4515e-15	3.32	A-B
8.187e-15	2.76	A-C
2.5078e-14	4.1	A-B
2.5526e-14	5.68	D-B



← Group pair for Max Mean Difference

↑ Maximum Difference between Group Means

Multiple Testing of Significance

- **Statistical problem**: Finding the differentially expressed genes measured simultaneously in the two or more groups of microarrays is the multiple test of significance problem, where many null hypotheses are tested simultaneously.

Procedures for Multiple Testing of Significance

Let a denote a pre-specified probability to reject the null-hypothesis for a given covariate. Let *m* gene tags measured simultaneously on a replicated microarray experiments

- The Bonferroni correction: If there are *m* null-hypotheses (tests), test each of these hypotheses to that level a/m . (very conservative: it dramatically increases the false-negative rate!)

- If m covariates are grouped in *j* families, than only *j* hypotheses should be tested at a significance level should be bigger~ ja/m

Hands-on Session 4

- Lab 9
- Total time: 15 minutes

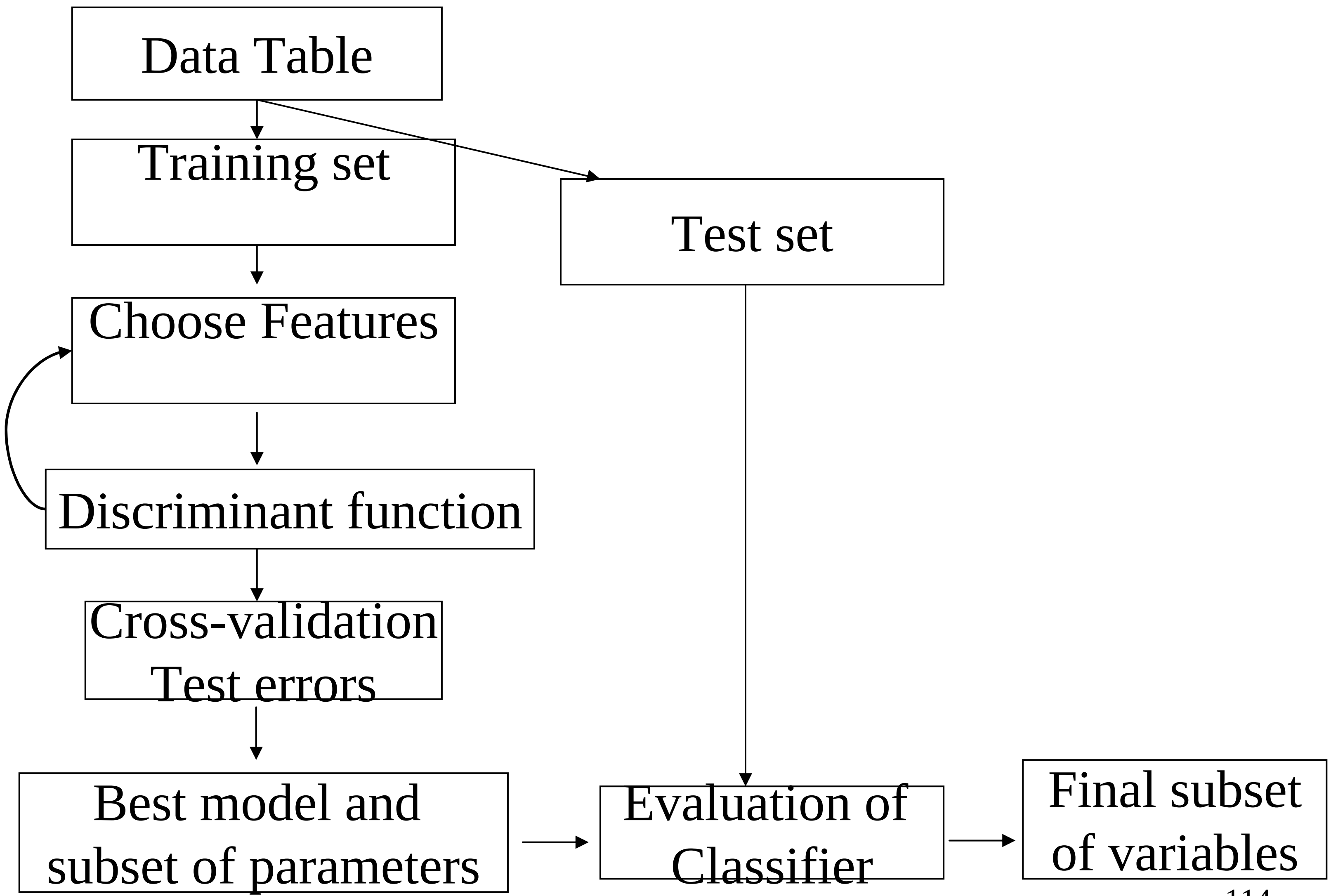
3. mADB dataset analysis tools

– Class Discovery: clustering, PCA, MDS
– Class Comparison: statistical analysis
– Class Prediction: PAM

Prediction Analysis for Microarrays (PAM): Class Prediction Supervised Model for Two or More Classes

- http://www-stat.stanford.edu/~tibs/PAM
- Provides a list of significant genes whose expression characterizes each class
- Estimates prediction error via cross-validation
- Imputes missing values in dataset

Design of the PAM algorithm



Calculating the Discriminant Function

For each gene, a centroid (a sample mean) is calculated for each given class.

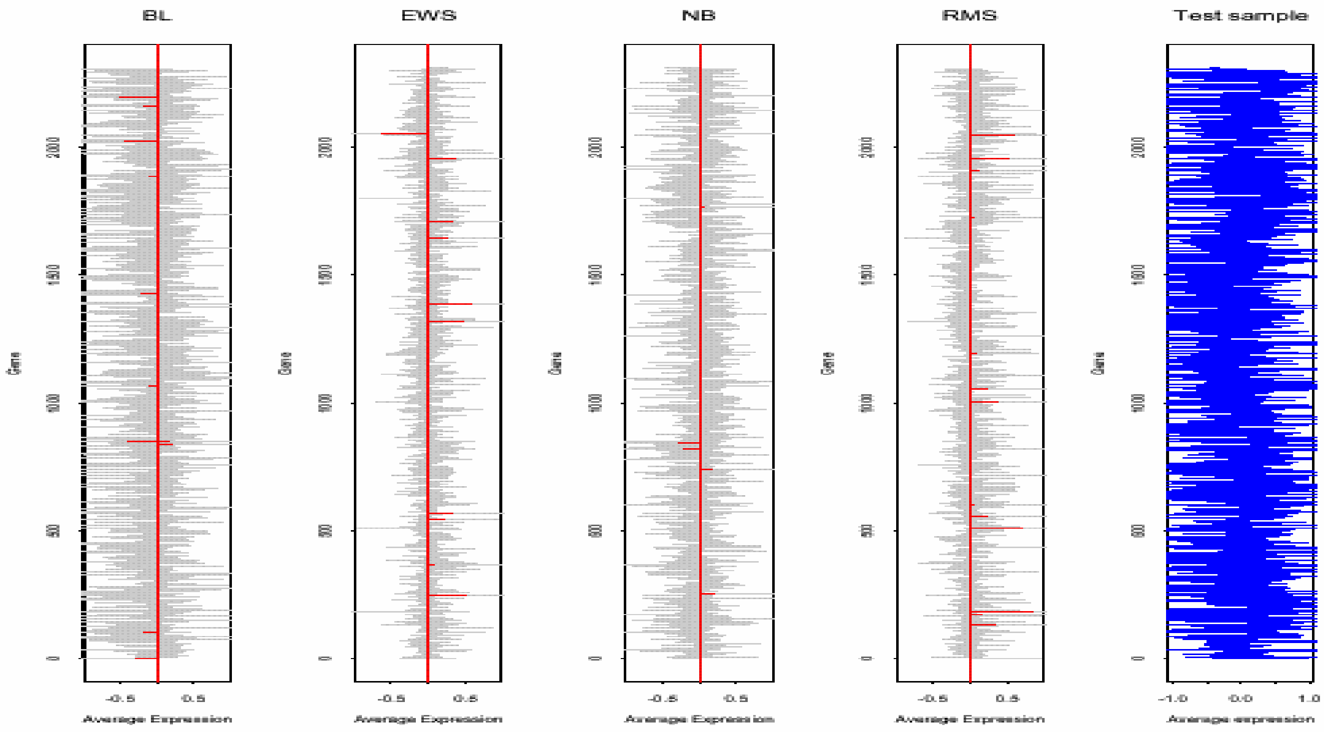
Briefly, the method computes a standardized centroid for each gene in each class. This is the average gene expression value in its class minus the overall gene expression average value divided by the standard deviation-like normalization factor for that gene.

centroid distance= (class avg – overall avg) / normalization factor.

Creates a normalized average gene expression profile for each class

Class Centroids

Class centroids



Classifying an Unknown Sample

A classifier takes the gene expression profile of a new sample (microarray) from test sets, and compares it to each of these class centroids. The class whose centroid that it is closest to, in squared distance, is the predicted class for that new sample.

.

K-fold Cross Validation

- The samples are divided up at random into K roughly equally sized parts.

Entire Data Set

50 Group A

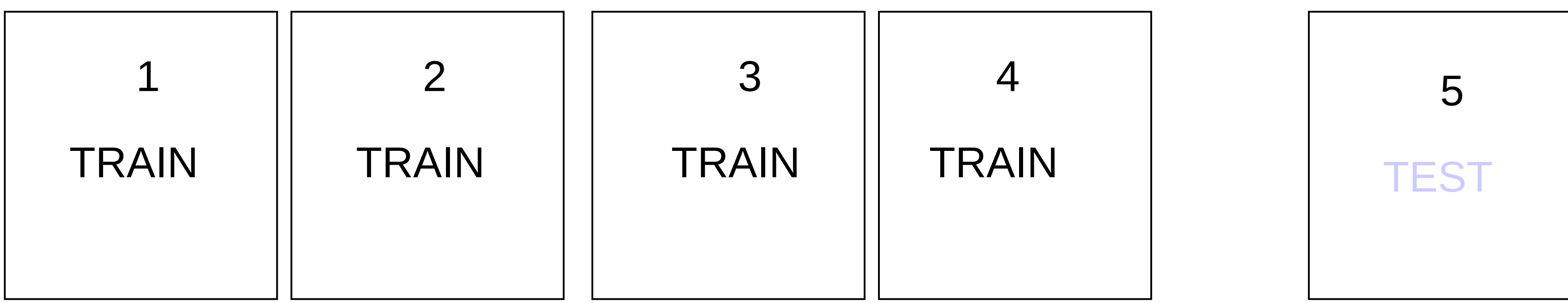
25 Group B

25 Group C

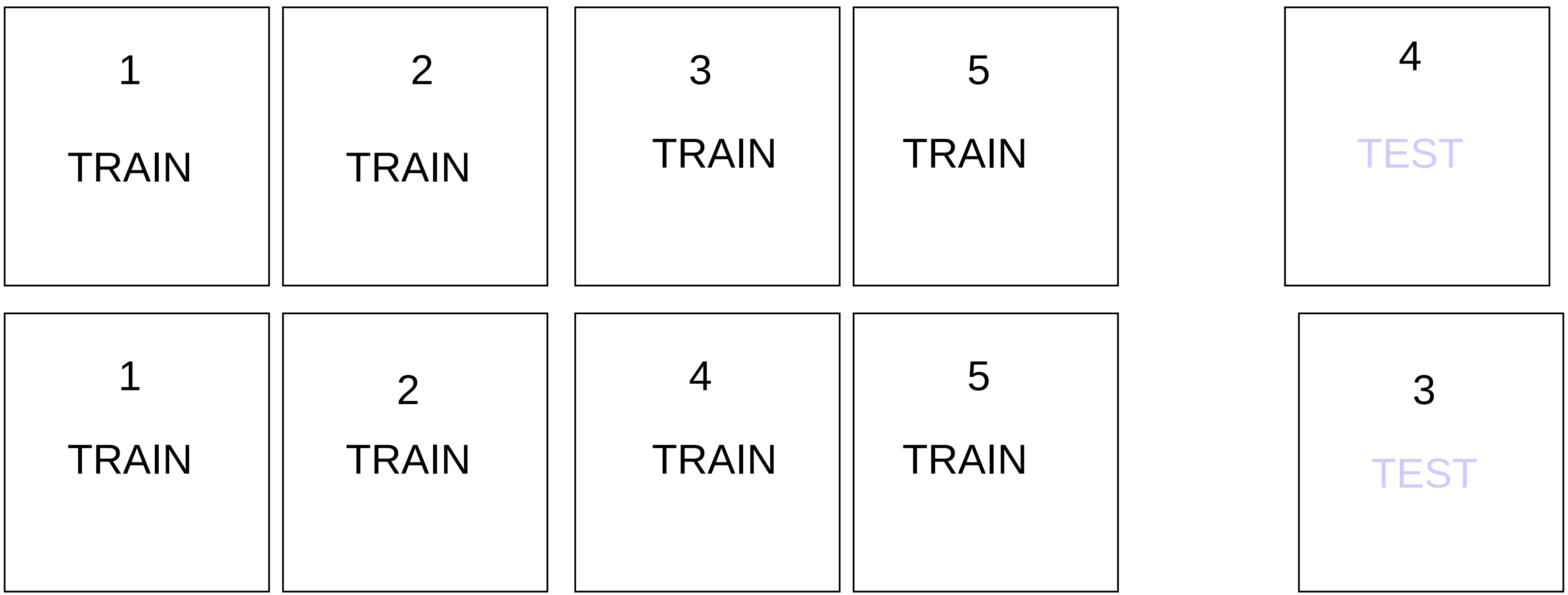
1	2	3	4	5
10 Group A	10 Group A	10 Group A	10 Group A	10 Group A
5 Group B	5 Group B	5 Group B	5 Group B	5 Group B
5 Group C	5 Group C	5 Group C	5 Group C	5 Group C

K-fold Cross Validation

For each part in turn, the classifier is built on the other K-1 parts then tested on the remaining part.



K-fold Cross Validation



etc….

Estimating Error Rate

PAM estimates a predicted error rate by averaging the error rate for each K cross validation

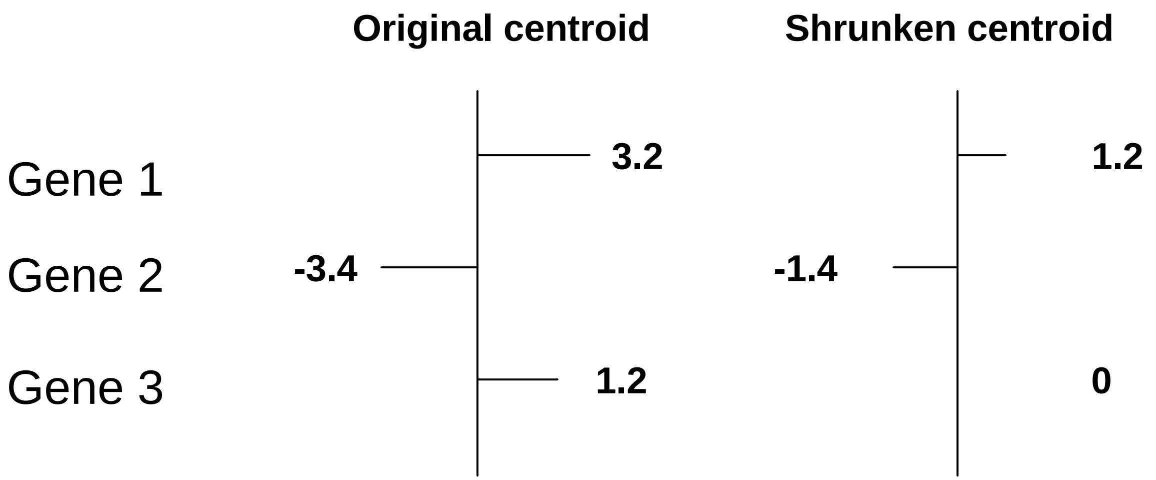
Reducing the feature set

Nearest shrunken centroid classification makes one important modification to standard nearest centroid classification. It "shrinks" each of the class centroids toward the overall centroid for all classes by an amount we call the threshold . This shrinkage consists of moving the centroid towards zero by threshold, setting it equal to zero if it hits zero.

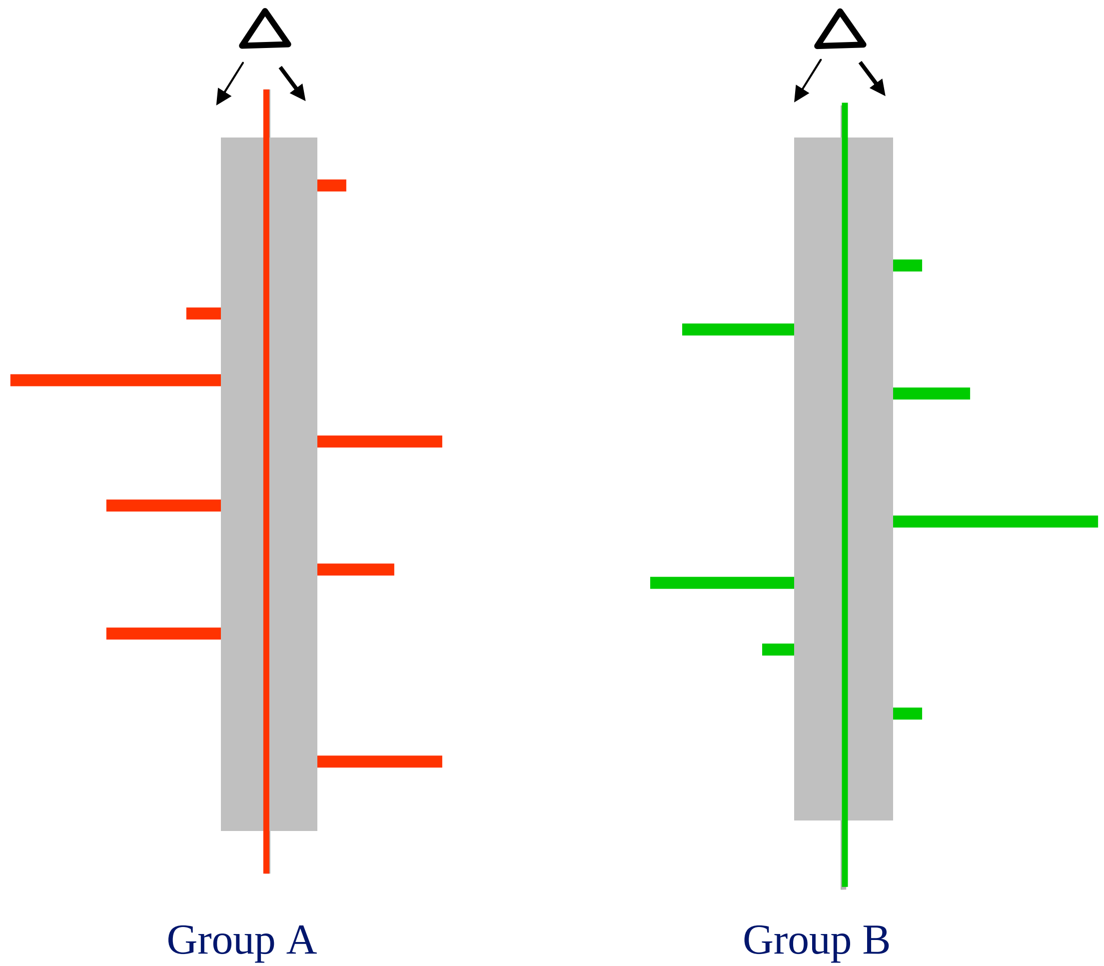
After shrinking the centroids, the new sample is classified by the usual nearest centroid rule, but using the shrunken class centroids.

Shrinking the centroid

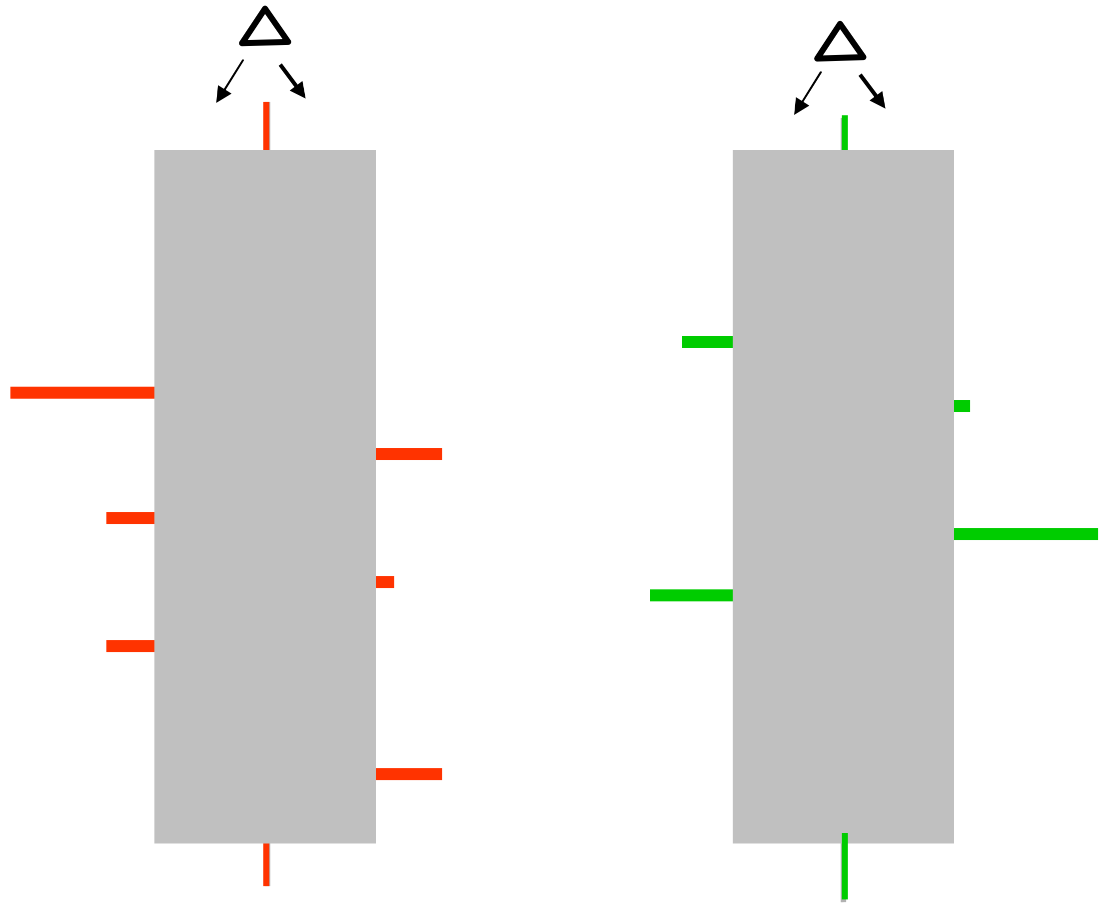
For example if threshold was 2.0, a centroid of 3.2 would be shrunk to 1.2, a centroid of -3.4 would be shrunk to -1.4, and a centroid of 1.2 would be shrunk to zero



Reduce Gene Number

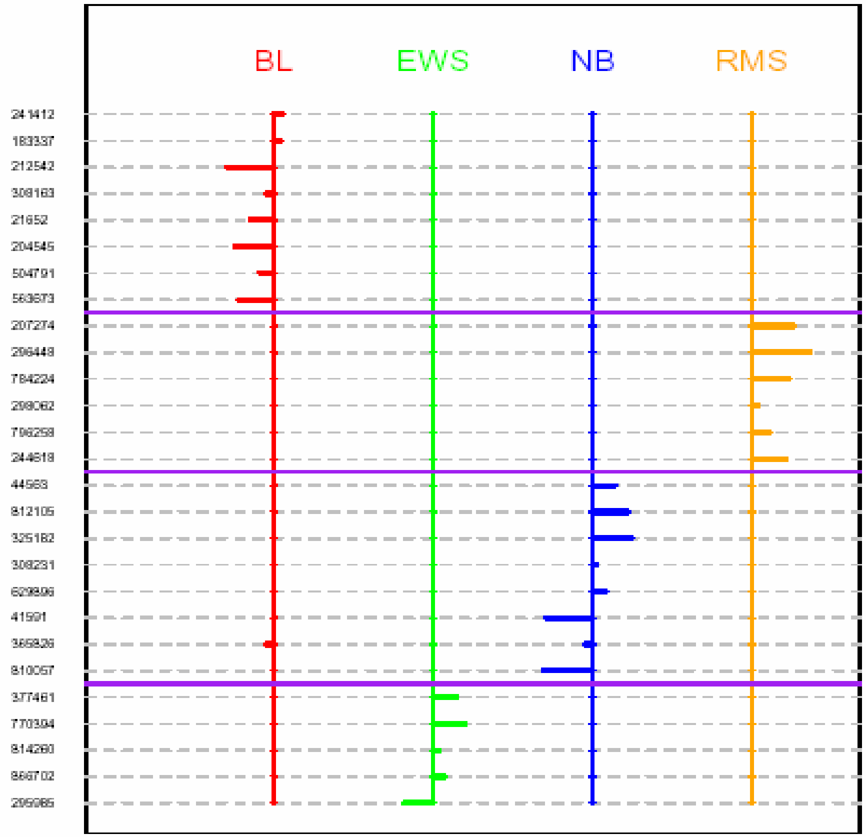


Incremental of threshold



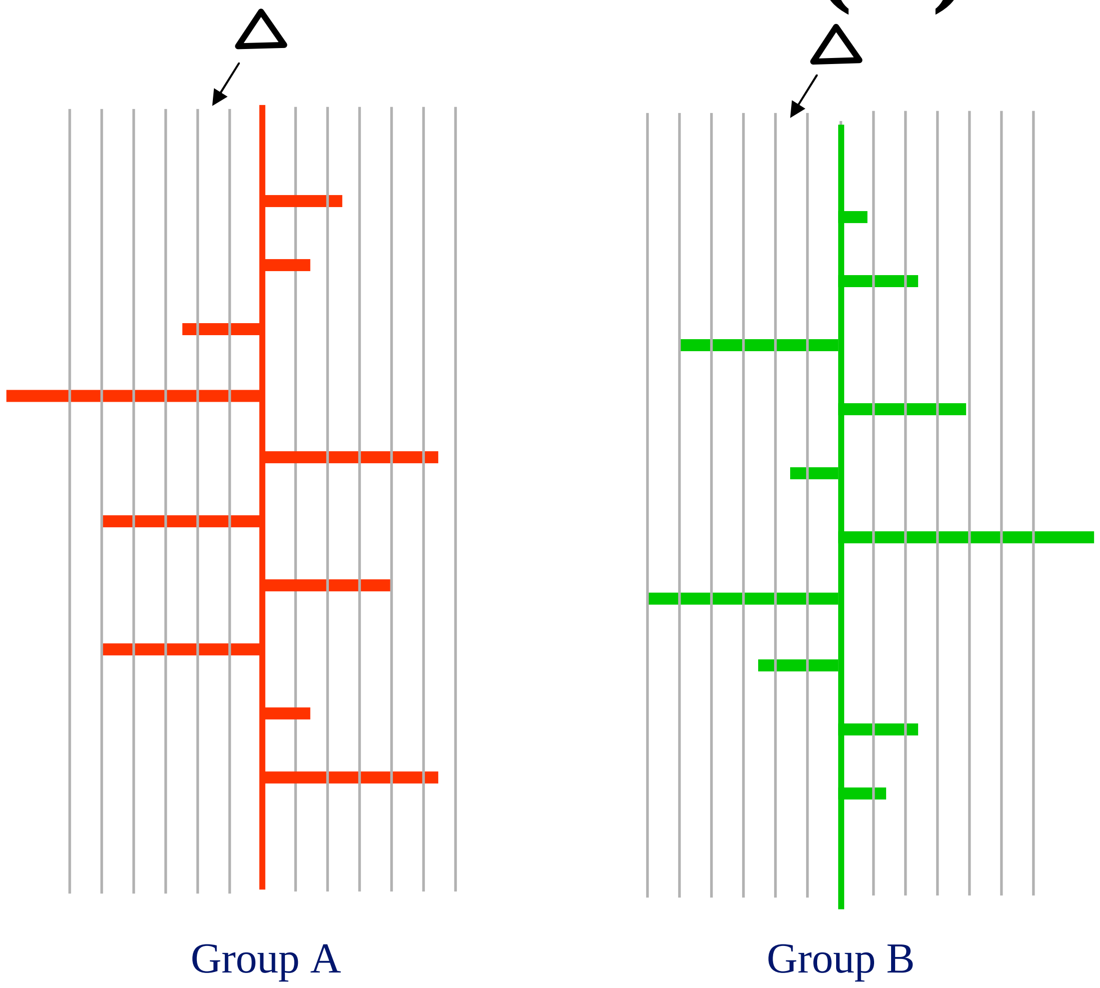
Group A   Group B

Prediction Model for SRBCT



- Compare model with new tumor tissues to make diagnosis

Multiple models with incremental threshold (△)



Misclassification Error

- Misclassification Error is calculated by averaging the errors from each of the cross validations.
- The model with lowest Misclassification Error is preferred.

Sample

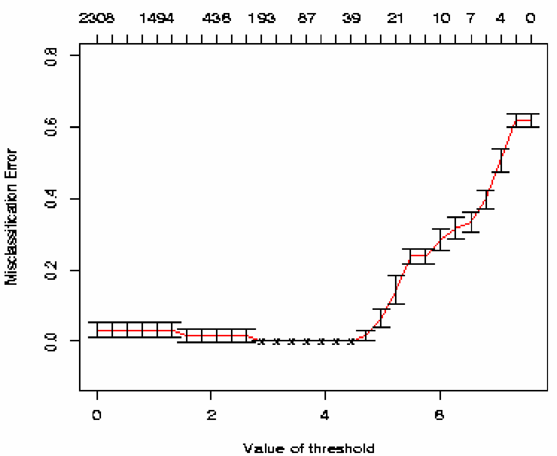
- 63 Arrays representing 4 groups
  - BL (Burkitt Lymphoma, n1=8)
  - EWS (Ewing, n2=23)
  - NB (neuroblastoma, n3=12)
  - RMS (rhabdomyosarcoma, n4=20)
- There are 2308 features (distinct gene probes)
- No missing values in array data sets
- Each group has an aggregate expression profile
- An unknown can be compared to each tumor class profile to predict which class it most likely belong

PAM Results

Clicking on a Delta value creates a new data Subset or enter
a Delta value at the bottom and Click "Create Subset".

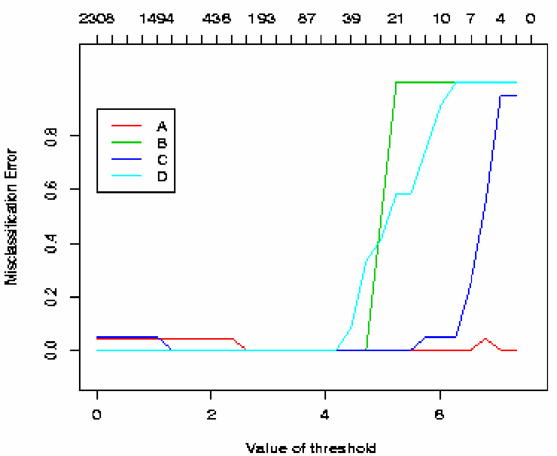
Shrinkage Delta	# of Genes	Misclass. Error
0.000	2308	0.032
0.262	2289	0.032
0.524	2145	0.032
0.786	1878	0.032
1.048	1494	0.032
1.309	1137	0.032
1.571	853	0.016
1.833	609	0.016
2.095	436	0.016
2.357	330	0.016
2.619	244	0.016
2.881 **	**193**	**0.000**
3.143 **	**151**	**0.000**
3.404 **	**107**	**0.000**
3.666 **	**87**	**0.000**
3.928 **	**68**	**0.000**
4.190 **	**52**	**0.000**
4.452 **	**39**	**0.000**
4.714	32	0.016
4.976	23	0.063
5.238	21	0.143
5.499	16	0.238
5.761	11	0.238
6.023	10	0.286
6.285	9	0.317
6.547	7	0.333
6.809	5	0.397
7.071	4	0.508

Create Subset



Misclassification error

Above as EPS, PDF, PNG



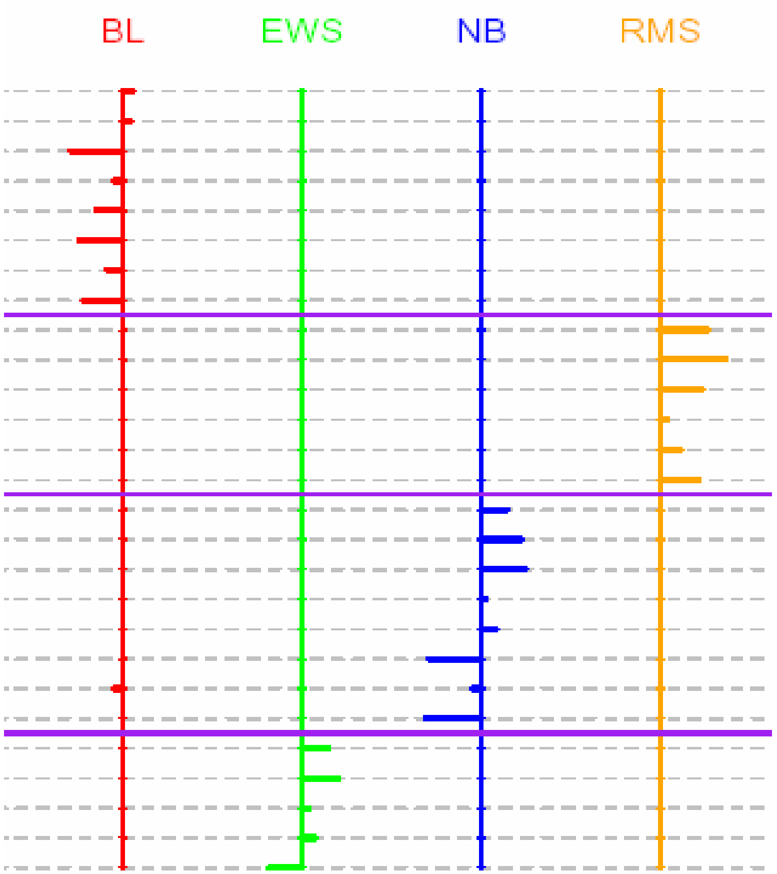
Link leads to the dataset with PAM model

Create new model by fill in a new Delta value

mAdb PAM Model

A Score	B Score	C Score	D Score
0.6092	-0.0866	0.0000	0.0000
0.0000	0.0000	0.0000	0.5862
-0.0696	0.0000	0.0000	0.5764
-0.5421	0.0000	0.0000	0.0000
0.5338	0.0000	0.0000	0.0000
0.0000	-0.5321	0.0000	0.0000
0.0000	0.0000	0.0000	0.4936
0.0000	-0.4873	0.0000	0.0000
0.0000	0.0000	0.0000	0.4821
0.0000	-0.4661	0.0000	0.0000
0.4380	0.0000	0.0000	0.0000
-0.0110	0.0000	0.0000	0.4269
0.0000	-0.4153	0.0000	0.0000
0.4086	0.0000	0.0000	0.0000
0.0000	0.0000	-0.3828	0.0000
0.3346	0.0000	0.0000	0.0000

=



PAM summary

- It generates models (classifiers) from microarray data with phenotype information
- It does automatic gene selection for each models.
- Misclassification errors are calculated with the data for model selection.
- Require adequate numbers of samples in each group

Hands-on Session 5

- Lab 10, Lab 11 (optional)
- Total time: 15 minutes

mAdb Development and Support Team:

- John Powell, Chief, BIMAS, CIT
- Liming Yang, Ph.D
- Jim Tomlin
- Xiaopeng Bian, Ph.D.**
- Esther Asaki*
- John Greene, Ph.D.*
- Vladimir Kuznetsov, Ph.D., Sci.D.*
- Kathleen Meyer*
- Tim Ruppert*

*SRA International contractor

** Postdoctoral Fellow

NCI Center for Cancer Research

Reducing the Burden of Cancer Through Exploration, Discovery and Translation

NIAID

SRA INTERNATIONAL, INC.

CIT

http://madb.nci.nih.gov
http://madb.niaid.nih.gov

For assistance, remember:

madb_support@bimas.cit.nih.gov

Determination of principal components is based on computing of eigen values and eigenvectors

Let n=2, and λ be an eigen value of matrix R

$$R = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$$

Basic model:

$$RV = \lambda V$$

where V is eigenvector

$$\begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$$

$$\mathrm{Det} \begin{pmatrix} 1-\lambda & r_{12} \\ r_{12} & 1-\lambda \end{pmatrix} = 0$$

$$\lambda^2 - 2\lambda + (1 - r_{12}) = 0$$

$$\lambda_1 = 1 + r_{12} \qquad \sigma^2(y_1) = \lambda_1$$

$$\lambda_2 = 1 - r_{12} \qquad \sigma^2(y_2) = \lambda_2$$

$$\lambda_1 + \lambda_2 = 2 = n$$

Examples:

$$r_{12} = 0.9 \qquad \lambda_1 = 1.9; \lambda_2 = 0.1$$

$$r_{12} = 1 \qquad \lambda_1 = 2; \lambda_2 = 0$$

$$r_{12} = 0 \qquad \lambda_1 = \lambda_2 = 1$$